Universidade de Brasília

IE – Departamento de Estatística

Análise de Diagnóstico para o Modelo de Regressão de Cox

Camila Farage de Gouveia

Orientadora: **Prof.ª Dra. Juliana Betini Fachini**

Brasília,

2013.

Camila Farage de Gouveia

Estágio Supervisionado 2

Análise de Diagnóstico para o Modelo de Regressão de Cox

Orientadora:

Prof.ª Dra.: **JULIANA BETINI FACHINI**

Trabalho de Conclusão de Curso apresentado à Universidade de Brasília, como exigência parcial para obtenção do título de bacharel em Estatística.

Brasília,

2013

AGRADECIMENTOS

À Professora Juliana Betini Fachini pela orientação nesse trabalho, por toda a paciência e dedicação.

Aos professores do Departamento de Estatística da Universidade de Brasília por todo o conhecimento passado a mim nesses anos na universidade, que possibilitou a realização desse trabalho.

À minha família por todo amor e apoio.

SUMÁRIO

1 INTRODUÇÃO

Este trabalho tem como intuito aplicar a análise de sobrevivência para analisar os dados de tempos de sobrevivência de pacientes submetidos a transplante de medula óssea para tratamento de leucemia mielóide crônica, no período de junho de 1986 a junho de 1998, no Centro de Transplante de Medula Óssea do Instituo Nacional do Câncer (Cemo – Inca).

A análise de sobrevivência é a metodologia adequada para modelar um estudo quando se tem como variável resposta o tempo até a ocorrência de um evento de interesse, sendo este tempo denominado tempo de falha.

Além deste tempo de falha, consideramos também o tempo de censura, que é o grande diferencial desta modelagem. O tempo de censura descreve os indivíduos que por algum motivo não concluíram o evento de interesse.

Outra característica presente em dados de sobrevivência são as covariáveis. Estas podem influenciar o tempo de sobrevivência ou de censura. Existem vários modelos de regressão que consideram essas características dos dados de sobrevivência, como o de Weibull e o de Cox. Neste trabalho será considerado o modelo de regressão de Cox, que, por possuir uma componente não paramétrica e não assumir uma distribuição de probabilidade aos dados é um modelo bastante flexível.

Após o ajuste do modelo, será feita uma análise de diagnóstico para validá-lo. Dessa forma, pode-se verificar se as suposições feitas são válidas e identificar possíveis características que possam influenciar as conclusões obtidas.

Toda a análise dos dados será feita com o auxílio do software R.

2 REVISÃO BIBLIOGRÁFICA

2.1 Análise de Sobrevivência

A metodologia de análise de sobrevivência é utilizada quando o objeto em estudo é o tempo desde algum instante pré-determinado até a ocorrência de um evento de interesse. Este tempo é definido como tempo de falha. Para uma boa análise deste tempo de falha, é necessário estabelecer padrões de escala e medida no início do estudo.

Diferentemente de outras técnicas de regressão, nessa análise conta-se com a presença de dados censurados, que são aqueles que por algum motivo não concluíram o evento de interesse. O tipo de censura de cada indivíduo no estudo pode ser à direita, quando ela ocorre após o início do estudo, à esquerda, quando ela ocorre antes do início do estudo, ou intervalar, quando não se conhece o tempo exato da censura, apenas o intervalo no qual ela acorreu. Neste trabalho será considerada apenas a censura à direita.

Sendo assim, a censura à direita pode ser dos seguintes tipos:

1. Tipo I: é estabelecido no início do estudo um tempo específico indicando o fim deste. Chegando a este tempo pré-estabelecido, os indivíduos que não concluíram o evento de interesse são censurados. A censura do tipo I pode ocorrer em duas situações:

   i. Quando todos os indivíduos iniciam o estudo ao mesmo tempo, sendo então censurados no mesmo instante τ. Neste caso, tem-se a censura simples do Tipo I. Os tempos são representados por:

   $$t_i = \min(T_i, \tau) \text{ e } \delta_i = \begin{cases} 1, \ se\ t_i = \ T_i \\ 0, se\ t_i = \ \tau. \end{cases}$$

   ii. Quando os indivíduos ingressam no estudo em tempos diferentes, de forma que eles também são censurados em tempos diferentes. O tempo observado para cada indivíduo é então definido por $L_1, L_2, \dots, L_n$, já que o tempo τ é contado a partir do instante que cada indivíduo ingressa no estudo. Neste caso, têm-se as censuras múltiplas do Tipo I e os tempos são definidos por:

   $$t_i = \min(T_i, L_i) \text{ e } \delta_i = \begin{cases} 1, \ se\ T_i \leq \ L_i \\ 0, se\ T_i > \ L_i. \end{cases}$$

2. Tipo II: é estabelecido no início do estudo um número máximo de falhas k. Dessa forma, todos os indivíduos que não concluíram o evento de interesse até essa k–ésima falha são censurados.
3. Tipo aleatória: algo não previsto ocorre e o indivíduo é obrigado a sair do estudo, ou até uma combinação dos primeiros dois tipos de censura.

O modelo determinado pela técnica de sobrevivência é então definido pelo tempo de sobrevivência *T*, que é uma variável aleatória não negativa e, geralmente, contínua. Conta-se com uma variável indicadora δ_i para determinar se o tempo é de falha ou censura. Sendo assim, cada indivíduo é representado pelo par (t_i, δ_i), de forma que δ_i é tal que,

$$\delta_i = \begin{cases} 1 \text{ se } t_i \text{ é tempo de falha,} \\ 0 \text{ se } t_i \text{ é tempo de censura.} \end{cases}$$

Usualmente, variáveis explicativas são incluídas no modelo, podendo estas influenciar o tempo de falha ou de censura. Estas covariáveis são representadas pelo vetor $\boldsymbol{x_i^T} = (x_{i1}, x_{i2}, \dots x_{ip})$ e cada indivíduo passa então a ser representado por $(t_i, \delta_i, \boldsymbol{x_i^T})$, onde $i = 1,2, \dots, n$ indica o número de indivíduos e $r = 1,2, \dots, p$ indica o número de covariáveis. É interessante verificar se essas variáveis são relacionadas entre si.

A variável T, que define o tempo de sobrevivência, pode ser representada pela função de densidade de probabilidade, $f(t)$, função de sobrevivência, $S(t)$, e função risco, $\lambda(t)$, sendo possível relacionar essas funções.

A função de densidade de probabilidade é dada como o limite da probabilidade de um indivíduo falhar em um intervalo de tempo $[t + \Delta t)$, por unidade de Δt (tamanho do intervalo), ou por unidade de tempo. Lee (1992) a definiu como:

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t}.$$

A função de sobrevivência, também conhecida como taxa de sobrevivência acumulada, é dada como a probabilidade de um indivíduo sobreviver a um tempo t. Ela é dada por:

$$S(t) = P(T > t) = \int_t^{\infty} f(x)dx.$$

$S(t)$ é uma função contínua monótona decrescente e possui as seguintes propriedades definidas por Lawless (2003):

1. $S(0) = 1$;
2. $S(\infty) = \lim_{t\to\infty} S(t) = 0$.

A função de risco, também chamada de taxa de falha, é definida como o limite da probabilidade de um indivíduo falhar no intervalor de tempo $[t, \Delta t)$, dado que este indivíduo sobreviveu até o tempo t, dividido pelo comprimento do intervalo. Essa função pode ser crescente, decrescente, constante ou não monótona. Lawless (2003) a definiu como:

$$\lambda(t) = \lim_{\Delta t\to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}.$$

Essa função também pode ser determinada em termos da função de densidade de probabilidade e da função de sobrevivência:

$$\lambda(t) = \frac{f(t)}{S(t)}.$$

A função risco é mais informativa que a função de sobrevivência, pois diferentes funções de sobrevivência podem possuir formas parecidas, enquanto que as taxas de falha podem diferir substancialmente.

Uma função importante é a função de risco acumulada, sendo representada por:

$$\Lambda(t) = \int_0^t \lambda(u)du.$$

Além da relação já explicitada da função de risco com a função de densidade de probabilidade e a função de sobrevivência, existem outras relações importantes, tais como:

1. $f(t) = \frac{\partial[1-S(t)]}{\partial t} = -S'(t)$;
2. $\lambda(t) = -\frac{S\prime(t)}{S(t)} = -\frac{\partial[logS(t)]}{\partial t}$;
3. $\log S(t) = -\int_0^t \lambda(u)du$.
4. $S(t) = \exp\{-\Lambda(t)\}$.

Tempos de sobrevivência são frequentemente aproximados por uma distribuição de probabilidade teórica, sendo possível, dessa forma, obter análises e informações mais completas e aprofundadas dos dados.

Existem diversas distribuições utilizadas na literatura, mas neste trabalho em específico, não será atribuída nenhuma distribuição de probabilidade para os tempos de sobrevivência, pois o modelo escolhido para representar os dados é o modelo de Cox, que conta com essa particularidade.

2.2 Modelo de Cox

O modelo de Cox é adequado quando a variável em estudo é o tempo decorrido até a ocorrência de um evento de interesse, podendo este tempo ser influenciado por covariáveis.

Determinado pela função de risco, o modelo de Cox é composto por um componente paramétrico e outro não paramétrico. Este último componente atribui grande versatilidade ao modelo, trazendo assim uma ampla utilização deste.

Também denominado de Modelo de Riscos Proporcionais, o modelo exige a suposição de que a razão das taxas de risco de dois indivíduos diferentes é constante no tempo. Dessa forma a razão não depende do tempo.

Para definir o modelo, considera-se uma função de risco basal, que define o componente não paramétrico, determinada $\lambda_0(t)$, e pelo componente paramétrico g($\boldsymbol{x'\beta}$), tal que $\boldsymbol{\beta}$ é o vetor de parâmetros associado às covariáveis, sendo estas determinadas pelo vetor $\boldsymbol{x}$. Comumente, este componente é definido como exp($\boldsymbol{x'\beta}$). Dessa forma, tem-se :

$$\lambda(t|\boldsymbol{x}) = \lambda_0(t)\, exp\big(x_1\beta_1 + x_2\beta_2 + \cdots + x_p\beta_p\big) = \\ = \lambda_0(t)\exp(\boldsymbol{x'\beta}), \tag{2}$$

em que $\lambda_0(t)$ é não negativa no tempo e não especificada, sendo chamada de função basal, pois, em grande parte dos modelos, quando $x = 0, \lambda_0(t) = \lambda(t)$.

Essa definição comprova a suposição de riscos proporcionais, já que:

$$\frac{\lambda_i(t|\boldsymbol{x_i})}{\lambda_j(t|\boldsymbol{x_j})} = \frac{\exp(\boldsymbol{x_i\beta})}{\exp(\boldsymbol{x_j\beta})} = k,$$

em que k é constante no tempo.

Existem outras fórmulas para a função $g(\boldsymbol{x'\beta})$ na literatura, mas a forma multiplicativa apresentada acima é a mais utilizada e será adotada no estudo em questão (Colosimo e Giolo, 2006).

Usualmente, estimam-se os parâmetros do modelo pelo método de máxima verossimilhança. O modelo de Cox, entretanto, por possuir um componente não paramétrico,

compromete este método de estimação. Sabe-se que a verossimilhança para dados censurados é dada por:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n}[f(t_i|\boldsymbol{x_i})]^{\delta i}\,[S(t_i|\boldsymbol{x_i})]^{1-\delta i} = \prod_{i=1}^{n}[\lambda(t_i|\boldsymbol{x_i})]^{\delta i}\,S(t_i|\boldsymbol{x_i}).$$

A função de sobrevivência associada pelo modelo de Cox definida em (2) é dada por:

$$S(t_i|\boldsymbol{x_i}) = \exp\left\{-\int_0^{t_i}\lambda_0(u)\,exp\{\boldsymbol{x'\beta}\}\,du\right\} = [S_0(t_i)]^{\exp\{\boldsymbol{x_i'\beta}\}}, \qquad (3)$$

de forma que a função de verossimilhança ao considerar (2) e (3) é escrita como:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n}[\lambda_0(t_i)\exp\{\boldsymbol{x_i'\beta}\}]^{\delta i}[S_0(t_i)]^{\exp\{\boldsymbol{x_i'\beta}\}},$$

que é função do componente não paramétrico.

Para contornar este problema, Cox (1975) propôs o método de máxima verossimilhança parcial. Condicionou a construção da função de verossimilhança ao histórico de falhas e censuras, eliminando assim esta função de perturbação da verossimilhança.

2.2.1 Função de Máxima Verossimilhança Parcial

Considera que uma amostra com n indivíduos possui número k de falhas distintas, tal que $k \leq n$, nos tempos $t1 < t2 < \ldots < tk$. Nesta metodologia empates não são considerados, pois a função de verossimilhança parcial supõe que os tempos são contínuos.

Primeiramente considera-se a probabilidade condicional da i-ésima observação vir a falhar no tempo ti, conhecendo quais observações estão sob risco em t_i. Colosimo e Giolo (2006), a partir da proposta de Cox (1975), definiram que:

$$P[indivíduo\ falhar\ em\ t_i \mid uma\ falha\ em\ ti\ e\ história\ até\ t_i] =$$

$$\frac{P[indivíduo\ falhar\ em\ t_i \mid sobreviveu\ a\ t_i\ e\ história\ até\ t_i]}{P[uma\ falha\ até\ t_i \mid história\ até\ t_i]} =$$

$$\frac{\boldsymbol{\lambda_i}(t \mid \boldsymbol{xi})}{\sum_{\boldsymbol{j}\,\epsilon\,\boldsymbol{R}(t_i)}\boldsymbol{\lambda_j}(t \mid \boldsymbol{x_j})} = \frac{\boldsymbol{\lambda_0}(t)\boldsymbol{exp}\{\boldsymbol{x_i'\beta}\}}{\sum_{\boldsymbol{j}\,\epsilon\,\boldsymbol{R}(t_i)}\boldsymbol{\lambda_0}(t)\,\boldsymbol{exp}\{\boldsymbol{x_j'\beta}\}} = \frac{\boldsymbol{exp}\{\boldsymbol{x_i'\beta}\}}{\sum_{\boldsymbol{j}\,\epsilon\,\boldsymbol{R}(t_i)}\boldsymbol{exp}\{\boldsymbol{x_j'\beta}\}}.$$

em que $R(t_i)$ é o conjunto dos índices das observações sob risco no tempo t_i.

Dessa maneira, elimina-se a função basal $\lambda_0(t)$. A função de verossimilhança passa a ser determinada pelo produtório dos termos acima, associados aos tempos de falha:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{k} \frac{\exp\{\boldsymbol{x_i}'\boldsymbol{\beta}\}}{\sum_{j \in R(t_i)} \exp\{\boldsymbol{x_j}'\boldsymbol{\beta}\}} = \prod_{i=1}^{n} \left(\frac{\exp\{\boldsymbol{x_i}'\boldsymbol{\beta}\}}{\sum_{j \in R(t_i)} \exp\{\boldsymbol{x_j}'\boldsymbol{\beta}\}} \right)^{\delta i},$$

sendo δ_i o indicador de falha ou censura.

Encontramos os valores de β que maximizam a função de verossimilhança parcial ao igualar a zero a derivada da função $l(\boldsymbol{\beta}) = \log\big(L(\boldsymbol{\beta})\big)$, isto é:

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \delta_i \left[x_i - \frac{\sum_{j \in R(t_i)} x_j exp\{\boldsymbol{x_i'}\widehat{\boldsymbol{\beta}}\}}{\sum_{j \in R(t_i)} \exp\{\boldsymbol{x_j'}\widehat{\boldsymbol{\beta}}\}} \right] = 0$$

Como dito anteriormente, a função de verossimilhança, ao supor que os tempos são contínuos, não inclui a ocorrência de empates. Isto não representa fielmente a realidade, já que dependendo da escala de medida utilizada em cada estudo, podem-se encontrar tempos de falha ou censura iguais para mais de um indivíduo.

Para adaptar a função de verossimilhança à realidade, Breslow (1972) e Peto (1972) propuseram uma aproximação. Seja s_i o vetor formado pelo somatório das correspondentes p covariáveis para os indivíduos que falharam no mesmo tempo t_i, tal que $i = 1, \ldots, k$. O número de falhas neste mesmo tempo será determinado por d_i. A partir disso, temos que a função de máxima verossimilhança parcial será aproximada por:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{k} \frac{\exp\{\boldsymbol{s_i}'\boldsymbol{\beta}\}}{[\sum_{j \in R(t_i)} \exp\{\boldsymbol{x_j}'\boldsymbol{\beta}\}]^{d_i}}$$

Essa aproximação deve ser usada quando não há um número muito grande de empates. Em situações que o número de empates em qualquer tempo é grande, o modelo de Cox para dados grupados é indicado (Lawless,1982, Prentice e Gloeckler, 1978).

Para analisar as estimativas dos parâmetros do modelo ajustado, podem-se construir intervalos de confiança e testar as hipóteses relativas aos coeficientes do modelo. Ao utilizar as propriedades assintóticas dos estimadores, ou seja:

$$\widehat{\boldsymbol{\beta}} \sim Np\left(\boldsymbol{\beta}, \Sigma(\widehat{\boldsymbol{\beta}})\right),$$

em que p é a dimensão de $\widehat{\boldsymbol{\beta}}$ e $\Sigma(\widehat{\boldsymbol{\beta}}) \approx -[E(\ddot{L}(\boldsymbol{\beta})]^{-1}$, dado que $\ddot{L}(\widehat{\boldsymbol{\beta}})$ é definida como a matriz de informação observada, dada por:

$$\ddot{L}(\widehat{\boldsymbol{\beta}}) = \frac{\partial^2 l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}.$$

Em casos onde essa esperança é difícil ou impossível de ser calculada, aproxima-se $\Sigma(\widehat{\boldsymbol{\beta}})$ à $[\ddot{L}(\boldsymbol{\beta})]^{-1}$, avaliada em $\beta = \hat{\beta}$.

Andersen e Gil (1982) apresentaram as provas mais gerais das propriedades destes estimadores, apesar de outros autores terem também estudados estas propriedades, como Cox (1975) e Tsiatis (1981).

2.2.2 Interpretação das Estimativas dos Coeficientes

A partir da função de taxa de falha determinada pelo modelo de Cox, observa-se que as covariáveis exercem o papel de acelerar ou desacelerar o risco. Com a razão das taxas de falhas, os coeficientes estimados podem ser interpretados, considerando a propriedade de riscos proporcionais.

Sabe-se que a razão da taxa de risco de dois indivíduos ao longo do estudo é constante e positiva. Se o objetivo é verificar uma covariável específica, tomam-se dois indivíduos com valores diferentes para esta covariável e valores iguais para as outras. A razão vai indicar a proporção em que o risco de um indivíduo é maior que a do outro considerando esta covariável. Dados os indivíduos i e j, e considerando a comparação pela covariável l, tem-se:

$$\frac{\lambda(t\,|\boldsymbol{x_i})}{\lambda(t\,|\boldsymbol{x_j})} = exp\{\boldsymbol{\beta_l}(\boldsymbol{x_{il}} - \boldsymbol{x_{jl}})\}. \qquad (4)$$

Para melhor compreensão, considere os dados utilizados neste trabalho. São 96 pacientes diagnosticados com leucemia mielóide crônica. A variável estudada é o tempo desde o transplante até o óbito do paciente. Suponha que x_l seja a covariável dicotômica indicando se os pacientes tiveram o desenvolvimento da doença de enxerto aguda, $x_l = 1$, ou não tiveram, $x_l = 0$. Infere-se a partir da função (4), que a taxa de risco dos pacientes que tiveram o desenvolvimento da doença é $exp\{\boldsymbol{\beta_l}\}$ vezes maior que a taxa de risco dos pacientes que não tiveram o desenvolvimento da doença, considerando fixas as outras covariáveis.

Agora, considerando a covariável fase da doença que apresenta três níveis: 0 se aguda, 1 se crônica e 2 se blástica. Considera-se a fase aguda como grupo controle. Neste caso, a comparação é realizada entre as taxas de risco do grupo controle e do grupo 1 e entre a taxa de risco do grupo controle e do grupo 2.

A interpretação muda quando há covariáveis contínuas no estudo. Dado que o parâmetro associado à covariável contínua é significativo, ao aumentar uma unidade na covariável, o risco de falha aumenta.

Hosmer e Lemeshow (1999) apresentam em seu trabalho maiores detalhes a respeito da interpretação das estimativas.

2.2.3 Estimativas para outras funções relacionadas à $\boldsymbol{\lambda_0(t)}$

Muitas vezes é desejado estimar a função de risco acumulado de base, assim como a função de sobrevivência de base. Estas estimativas são utilizadas, principalmente, para auxiliar na verificação da adequação do ajuste do modelo.

Sabe-se que $\lambda_0(t)$ não é determinada de forma paramétrica. Se fosse, seria possível estimá-la a partir do método de máxima verossimilhança. Como o método de máxima verossimilhança parcial, que é utilizado no modelo de Cox, absorve esta função, encontrou-se uma forma não paramétrica para estimá-la.

Breslow (1972) propôs, para a estimativa de $\Lambda_o(t)$, uma função escada com saltos nos tempos distintos de falha, definida como:

$$\widehat{\Lambda}_0(t) = \sum_{j:\, t_j<t} \frac{d_j}{\sum_{l\,\epsilon\, R_j} \exp\{\boldsymbol{x'}_l\widehat{\boldsymbol{\beta}}\}},$$

em que d_j é o número de falhas em t_j. A partir dessa equação, determinam-se as estimativas das funções $S_0(t)$ e $S(t|x)$. Tem-se:

$$\hat{S}_0(t) = exp\{-\widehat{\Lambda}_0(t)\},$$

$$\hat{S}(t|\boldsymbol{x}) = [\hat{S}_0(t)^{exp\{\boldsymbol{x'}\widehat{\boldsymbol{\beta}}\}}],$$

Observa-se que ambas são funções escada decrescentes com o tempo e que na ausência de covariáveis, a estimativa para $\widehat{\Lambda}_0(t)$ se reduz ao estimador de Nelson-Aalen. Por este motivo, este estimador proposto por Breslow também é chamado de estimador de Nelson-Aalen-Breslow.

2.3 Análise de Diagnóstico

Após ajustar os dados de um estudo a um modelo específico, deve-se verificar se o modelo escolhido é o mais adequado, se este de fato descreve o comportamento das observações e se as suposições necessárias para o seu uso são válidas. Caso o ajuste não seja o mais apropriado, conclusões equivocadas podem ser obtidas.

Sendo assim, a análise de diagnóstico é imprescindível para a validação de qualquer modelo. Através dela, podem-se encontrar observações influentes nos dados, isto é, observações que se omitidas no estudo ou se perturbadas podem trazer mudanças na análise estatística. Essas observações devem ser avaliadas individualmente para que seja decidido se estas devem ser mantidas ou não.

As metodologias utilizadas para obter tal análise são: Análise de Resíduos e Influência Global. A primeira analisa o comportamento dos resíduos do ajuste, para que seja possível verificar se as suposições exigidas para o uso do modelo são validas e se existem dados discrepantes na amostra. Em dados de sobrevivência, os resíduos usualmente utilizados são o *Martingal* e o *Deviance*.

A segunda metodologia constitui na deleção de casos, isto é, retirar o i-ésimo indivíduo do estudo, para assim averiguar se houve alguma mudança importante no modelo. Para julgar se cada observação é influente ou não, utiliza-se a Distância de Cook, proposta por Cook (1977), que mede a influência da retirada de cada indivíduo nas estimativas dos parâmetros em modelos de regressão linear. Outra medida usada é o Afastamento da Verossimilhança, que mede a influência de cada indivíduo na verossimilhança.

As metodologias de Análise de Influência Global são apresentadas em Gomes (2007), Fachini (2006) e Fachini (2011).

2.3.1 Análise de Resíduos

Após ajustar um modelo, deve-se avaliar se este foi bem escolhido para os dados em estudo, se as suposições necessárias são válidas e se se existem pontos discrepantes ou mal ajustados.

Sendo assim, essa análise será dividida em duas etapas:

i. Verificação da suposição de riscos proporcionais

Utiliza-se um método gráfico para analisar essa suposição a partir dos resíduos de Schoenfeld (1982). Dado que o i-ésimo indivíduo falhou, considerando um vetor de covariáveis $x_i = (x_{i1}, x_{12}, \dots, x_{ip})'$, o vetor de resíduos de Schoenfeld é $r_i = (r_{i1}, r_{i2}, \dots, r_{ip})$, em que r_{iq}, para o vetor de covariáveis $q = 1, \dots, p$, é definido por:

$$r_{iq} = x_{iq} - \frac{\sum_{j \in R(t_i)} x_{jq} \exp\{\boldsymbol{x}_j'\widehat{\boldsymbol{\beta}}\}}{\sum_{j \in R(t_i)} \exp\{\boldsymbol{x}_j'\widehat{\boldsymbol{\beta}}\}},$$

de forma que os resíduos são definidos em cada falha e que $\sum_i r_i = o$. Para que seja possível considerar a estrutura de correlação dos resíduos, comumente utiliza-se a forma padronizada desses, definida por:

$$s_i^* = \left[\ddot{L}(\widehat{\boldsymbol{\beta}})\right]^{-1} r_i,$$

em que $\ddot{L}(\widehat{\boldsymbol{\beta}})$ é a matriz de informação observada.

Analisam-se esses resíduos construindo um gráfico de $\boldsymbol{\beta}_{q(t)}$ versus $g(t) = t$. Se os riscos forem proporcionais, espera-se encontrar uma reta horizontal neste gráfico, isto é, inclinação próxima à zero sugere evidências de que os riscos de fato são proporcionais, de forma que os resíduos se comportem de forma aleatória ao redor desta reta.

A interpretação deste gráfico pode ser muito subjetiva e, por isso, definiram-se medidas estatísticas para que fosse possível obter conclusões mais concisas.

Uma das maneiras para testar a hipótese de riscos proporcionais, é utilizar o coeficiente de correlação de Pearson (ρ) entre os resíduos padronizados de Schoenfeld e g(t), para cada covariável. Para valores de ρ próximos à zero, não há evidências para rejeitar a hipótese nula de riscos proporcionais.

Para poder verificar esta hipótese para todas as covariáveis no modelo, assume-se que $g_{q(t)} = g(t)$, e utiliza-se a seguinte estatística de teste:

$$T = \frac{(g - \bar{g})' S^* \ddot{L} S^{*\prime} (g - \bar{g})}{d \sum_k (g_k - \bar{g})^2},$$

em que d é o número de falhas e $S^* = d\boldsymbol{R}\ddot{L}^{-1}$, tal que $\boldsymbol{R}$ é a matriz $d \times p$ dos resíduos de Schoenfeld não-padronizados. Sob a hipótese nula de proporcionalidade dos riscos, T tem aproximadamente distribuição qui-quadrado com p graus de liberdade. Há evidências para rejeitar a hipótese para valores de $T > \chi^2_{p,1-\alpha}$.

ii. Verificação da qualidade geral do modelo

Existem diversos resíduos utilizados para tal propósito. Comumente, em dados de sobrevivência opta-se pelo uso dos resíduos *Martingal* e *Deviance.*

O resíduo *Martingal*, que serve para apontar possíveis observações influentes, é definido por:

$$\widehat{M}_i = \delta_i - \widehat{\Lambda}_0(t)\exp\{\sum_{k=1}^{p} x_{ik}\widehat{\boldsymbol{\beta}}_{\boldsymbol{k}}\},$$

em que o segundo termo da equação à direita é o resíduo de Cox-Snell do modelo de Cox.

Com o intuito de tornar este resíduo simétrico com média zero e variância 1, realiza-se uma transformação, de forma que o resíduo transformado passa então a ser chamado de resíduo *Deviance*, dado por:

$$\widehat{Dev} = sinal(\widehat{M}_i)\{-2[\widehat{M}_i + \delta_i \log(\delta_i - \widehat{M}_i)]\}^{1/2}.$$

Se o modelo ajustado for apropriado, os resíduos devem apresentar um comportamento aleatório em torno de zero. Assim sendo, a partir de uma análise gráfica é possível avaliar a adequação do modelo e identificar pontos discrepantes.

2.3.2 Influência Global

A partir da deleção de casos, verifica-se se a retirada de determinadas observações trazem mudanças expressivas no modelo. Dessa forma é possível identificar observações que possuam grande influência e que devem ser analisadas com uma maior atenção.

Após a retirada de cada observação separadamente, utiliza-se as medidas Distância de Cook Generalizada e Afastamento da Verossimilhança para quantificar essa influência no modelo.

Ao retirar a observação i, a estimativa de máxima verossimilhança de um parâmetro $\boldsymbol{\beta}$ torna-se $\widehat{\boldsymbol{\beta}}_{(\boldsymbol{i})}$. Com a diferença entre as estimativas de $\widehat{\boldsymbol{\beta}}$ e $\widehat{\boldsymbol{\beta}}_{(\boldsymbol{i})}$, calculam-se as medidas mencionadas.

A distância de Cook Generalizada é dada por:

$$GD_i = \left(\widehat{\boldsymbol{\beta}}_{(\boldsymbol{i})} - \widehat{\boldsymbol{\beta}}\right)^T M\left(\widehat{\boldsymbol{\beta}}_{(\boldsymbol{i})} - \widehat{\boldsymbol{\beta}}\right),$$

em que M pode ser definida de diversas maneiras. Usualmente, utiliza-se $M = -\ddot{L}\left(\widehat{\boldsymbol{\beta}}\right)$ ou $M = \left[-\ddot{L}\left(\widehat{\boldsymbol{\beta}}\right)\right]^{-1}$. O termo $\ddot{L}\left(\widehat{\boldsymbol{\beta}}\right)$ se refere à matriz de informação observada, tal que:

$$\ddot{L}\left(\widehat{\boldsymbol{\beta}}\right) = \frac{\partial^2 l\left(\widehat{\boldsymbol{\beta}}\right)}{\partial\widehat{\boldsymbol{\beta}}\partial\widehat{\boldsymbol{\beta}}^T}.$$

Por sua vez, o Afastamento da Verossimilhança é calculado a partir do logaritmo da função de máxima verossimilhança $l(\boldsymbol{\beta})$ e $l(\boldsymbol{\beta}_{(\boldsymbol{i})})$:

$$LD_i = 2\left[\, l\left(\widehat{\boldsymbol{\beta}}\right) - \boldsymbol{l}\left(\widehat{\boldsymbol{\beta}}_{(\boldsymbol{i})}\right)\right].$$

Caso o valor de alguma dessas medidas seja alto, deve-se avaliar se a observação em estudo compromete as estimativas do modelo, podendo também indicar um dado discrepante ou erro na coleta dos dados.

Para melhor avaliar essas informações, são construídos gráficos das medidas contra a ordem das observações. Isso possibilita a identificação de possíveis observações influentes, dado que essas ficam mais distantes das outras observações.

2.3.3 Impacto das Observações Influentes

Para quantificar a influência que cada observação detectada pode trazer ao ajuste, definiu-se a mudança relativa do modelo como:

$$RC = [\widehat{\theta_J} - \widehat{\theta}_{j(i)}]/\widehat{\theta_J},$$

em que *I* se refere ao conjunto de observações retiradas da amostra.

Quanto maior a mudança relativa, maior é a influência da observação no modelo.

3 RESULTADOS E DISCUSSÕES

Neste trabalho serão analisados dados provenientes de uma coorte de 96 pacientes submetidos a transplante de medula óssea para tratamento de leucemia mielóide crônica, no período de junho de 1986 a junho de 1998, no Centro de Transplante de Medula Óssea do Instituo Nacional do Câncer (Cemo – Inca).

O objetivo do estudo foi analisar os efeitos de fatores prognósticos para a ocorrência de doença do enxerto contra o hospedeiro aguda e crônica, da sobrevivência livre de doença e da sobrevivência global.

As variáveis em estudo são:

- $\boldsymbol{T}$: variável resposta; tempo de sobrevivência de cada paciente; indica os dias desde o início do tratamento até o óbito do indivíduo ou até a censura deste;
- $\boldsymbol{\delta}$: variável indicadora de censura; informa se o tempo é de censura ou de falha ($0 = não, 1 = sim$);
- $\boldsymbol{X_1}$: sexo do indivíduo ($0 = masculinho, 1 = feminino$);
- $\boldsymbol{X_2}$**:** idade na data do transplante;
- $\boldsymbol{X_3}$: indica se houve ou não recuperação de plaquetas ($0 = não, 1 = sim$);
- $\boldsymbol{X_4}$: dias até a recuperação de plaquetas;
- $\boldsymbol{X_5}$: indica se houve o desenvolvimento da doença de enxerto aguda ($0 = não, 1 = sim$);
- $\boldsymbol{X_6}$: tempo até o desenvolvimento da doença enxerto aguda;
- $\boldsymbol{X_7}$: indica se houve o desenvolvimento da doença enxerto crônica ($0 = não, 1 = sim$);
- $\boldsymbol{X_8}$: tempo até o desenvolvimento da doença enxerto crônica;
- $\boldsymbol{X_9}$: fase da doença na data do transplante ($1 = crônica, 2 = aguda, 3 = blástica$).

3.1 Análise Descritiva

Para compreender melhor como os dados se comportam, de forma que a modelagem corresponda à realidade desses, foi feita uma análise exploratória dos dados, obtendo-se os seguintes resultados:

Tabela 1: Medidas descritivas considerando a variável resposta tempo de sobrevivência dos pacientes submetidos ao transplante de medula óssea no Cemo-Inca.

Mediana dos Tempos de Sobrevivência	% Censura Global	% Censura dado que $X_3 = 1$	% Censura dado que $X_9 = 1$	% Censura dado que $X_9 = 2$	% Censura dado que $X_9 = 3$
70,5	48,96%	58,97%	58,57%	30,00%	0,00%

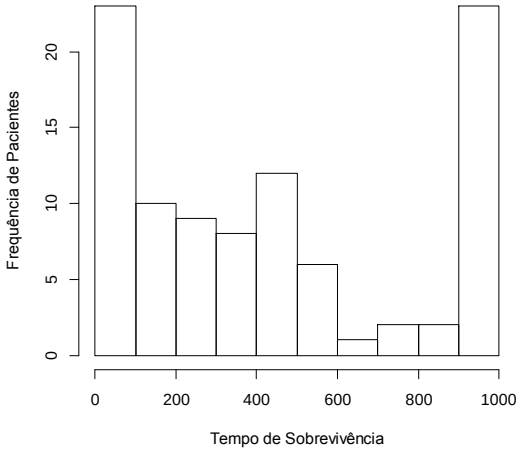

Figura 1: Histograma dos tempos de sobrevivência dos pacientes submetidos ao transplante de medula óssea no Cemo-Inca.

Na Tabela 1, observa-se o valor 370,5 da mediana e, dado que o estudo termina no tempo 1000, conclui-se que no início os indivíduos morrem mais rapidamente. Também pode-se observar os valores das porcentagens de censura em cada grupo considerando as covariáveis.

A Figura 1 mostra que, a princípio, a frequência dos tempos de sobrevivência diminui, mas que ao final do estudo (tempo=1000) essa frequência aumenta bastante. Isso é um indicativo de indivíduos censurados.

Os dados da amostra foram analisados, primeiramente não considerando as covariáveis, através do método de Kaplan-Meier para estimar a curva de sobrevivência dos pacientes. Obtiveram-se os seguintes gráficos:

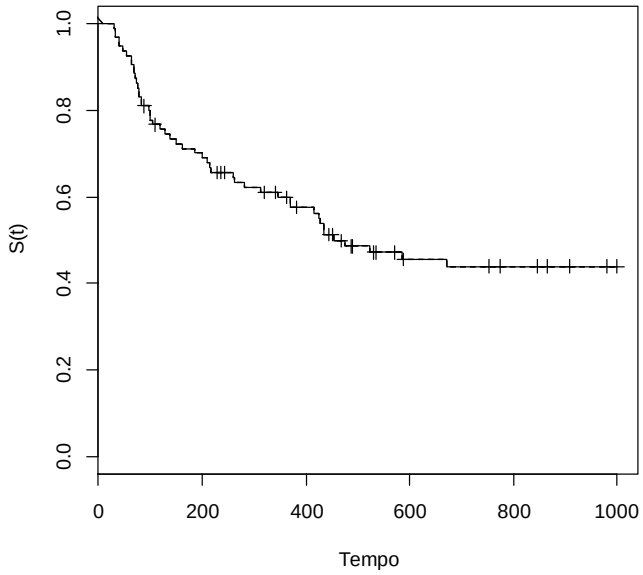

Figura 2: Gráfico da função de sobrevivência dos pacientes submetidos ao transplante de medula óssea no Cemo-Inca, estimada através do método de Kaplan-Meier.

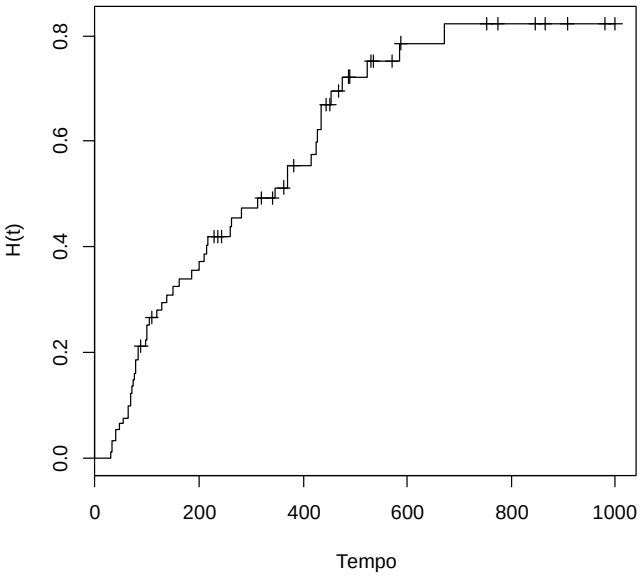

Figura 3: Gráfico da função de risco acumulado dos pacientes submetidos ao transplante de medula óssea no Cemo-Inca, estimado através do método de Kaplan-Meier.

Verificou-se pela Figura 2 que a curva de sobrevivência fica constante próxima ao valor 0,4, pouco depois de 600 dias. A curva não tende à zero, o que é um indicativo de que possam ter indivíduos curados nos dados. Para verificar a presença destes indivíduos nos dados, modelos com fração de cura podem ser utilizados. Neste trabalho, a princípio, não serão

considerados estes modelos. Ao analisar a Figura 3, verifica-se que a função de risco acumulada fica constante próxima ao valor 0,8.

Após essa análise prévia, tratou-se dos dados considerando cada covariável:

- $\boldsymbol{X_1}$**:** Sexo

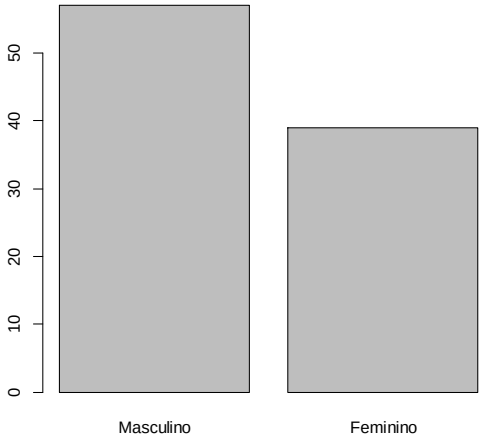

Figura 4a: Gráfico da frequência dos pacientes do sexo masculino e feminino.

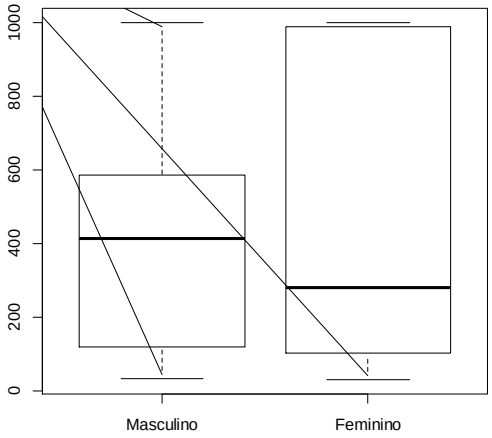

Figura 4b: Boxplot dos tempos de sobrevivência do grupo de pacientes do sexo masculino e feminino.

A Figura 4a mostra que há um maior número de pacientes do sexo masculino do que feminino. Ao construir o boxplot (Figura 4b), verifica-se que o comportamento dos tempos de sobrevivência dos dois grupos de pacientes definidos pelo sexo é similar.

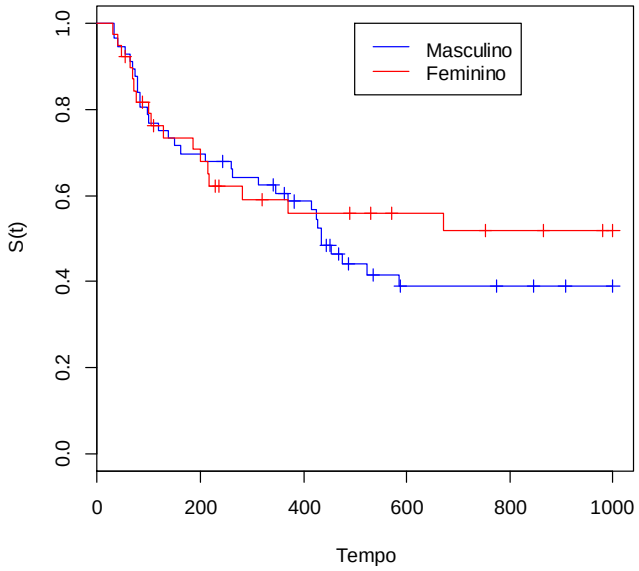

Figura 5: Gráfico da função de sobrevivência, considerando a covariável X_1 estimada pelo método de Kaplan Meier.

Avaliando pela Figura 5 as duas curvas de sobrevivência estimada por Kaplan-Meyer, observou-se que elas são bem próximas no início, mas que ao passar do tempo elas se distanciam um pouco, ficando constantes com valores diferentes.

A partir do teste de Wilcoxon, verificou-se a hipótese nula de que as curvas de sobrevivência não diferem para os dois sexos. Esse teste foi utilizado ao invés do teste de Log-Rank, pois este último exige que os riscos sejam proporcionais, o que não é comprovado para estes dados. Com um p-valor de 0.673, dado um nível de significância de 5%, não se rejeita a hipótese nula, havendo assim evidências para considerar que as curvas de sobrevivência dos pacientes não diferem para o sexo masculino e feminino. Essa conclusão confirma o observado a partir da Figura 4b, o boxplot dos dois grupos.

- $\boldsymbol{X_2}$: Idade

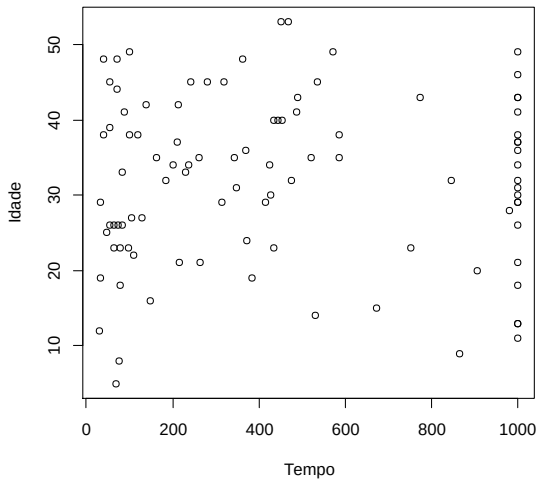

Figura 6a: Gráfico de dispersão da idade dos pacientes ao longo do tempo.

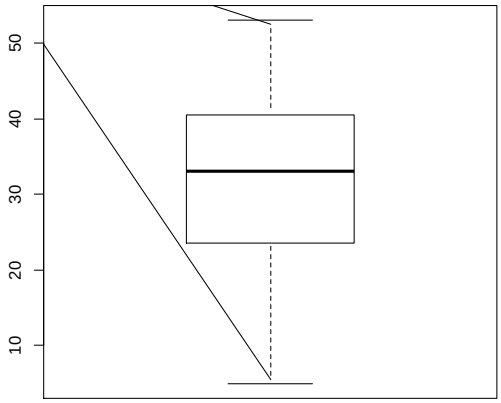

Figura 6b: Boxplot da idade dos pacientes.

A Figura 6a e a Figura 6b ilustram o maior indício de falhas no início do estudo e que os pacientes em sua maioria têm entre 20 e 50 anos. Existe uma predominância de observações no tempo 1000, o que é explicado pela censurada desses indivíduos ao final do estudo. Observa-se uma correlação fraca entre os tempos de sobrevivência e a idade dos pacientes a partir da Figura 6a.

- X_3: Recuperação de Plaquetas

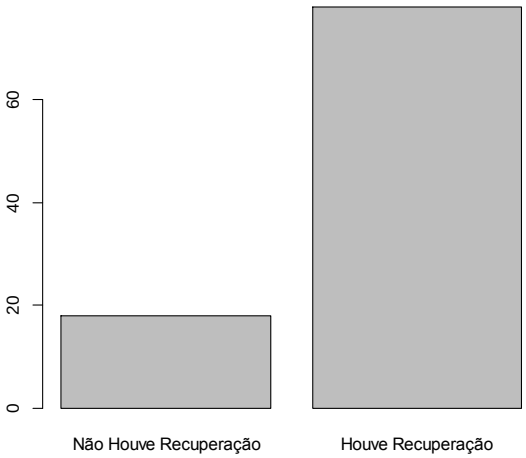

Figura 7a: Gráfico da frequência dos pacientes de cada grupo considerando a recuperação de plaquetas.

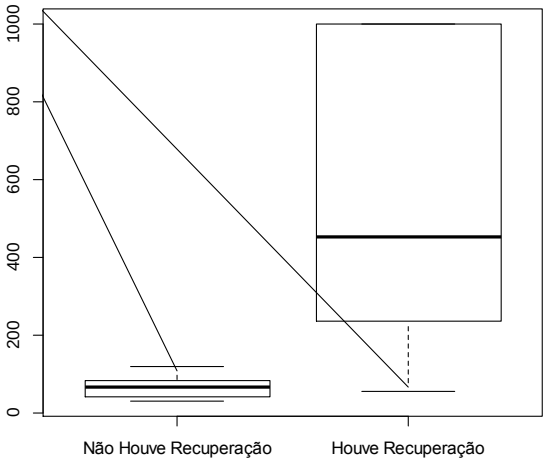

Figura 7b: Boxplot dos tempos de sobrevivência de cada grupo considerando a recuperação de plaquetas.

A Figura 7a evidencia uma maior predominância de indivíduos que tiveram recuperação de plaquetas. O boxplot ilustra também a maior amplitude dos tempos de sobrevivência dos indivíduos que tiveram essa recuperação e indica que então há diferença entre os tempos de sobrevivência dos pacientes considerando essa variável.

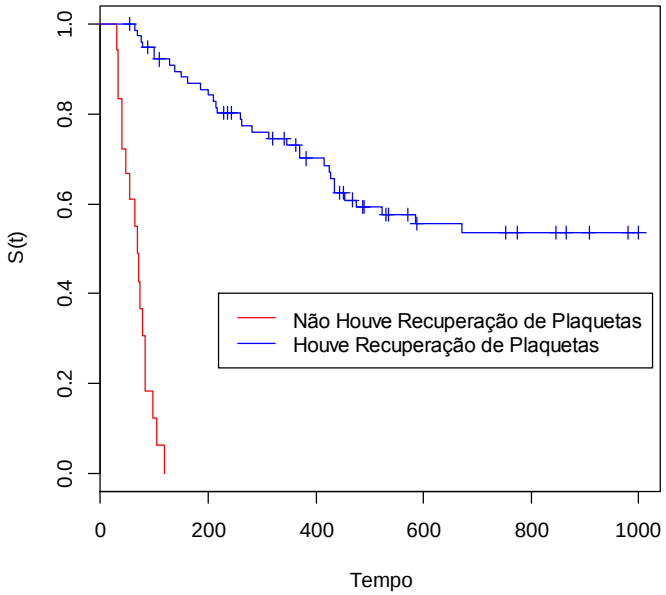

Figura 8: Gráfico da função de sobrevivência dos pacientes, considerando a variável X_3, recuperação de plaquetas, estimada pelo método de Kaplan-Meier.

Para testar se a variável X_3 influencia o tempo de sobrevivência dos pacientes foi realizado o teste de Log-Rank, já que se pode assumir que os riscos são proporcionais dado que as curvas de sobrevivência estimadas para os dois grupos não se cruzam. Com p-valor<0.0001, rejeitou-se a hipótese nula, tendo assim evidências para acreditar que as curvas de sobrevivência dos pacientes que tiveram recuperação de plaquetas e dos que não tiveram são diferentes. Neste caso, calculou-se também o risco relativo e concluiu-se que os pacientes que tiveram recuperação de plaquetas têm, aproximadamente, 11.29 mais chances de sobreviver que os outros pacientes. Esse resultado confirma a conclusão preliminar a partir da Figura 7a e da Figura 7b.

- X_4: Tempo de Recuperação de Plaquetas

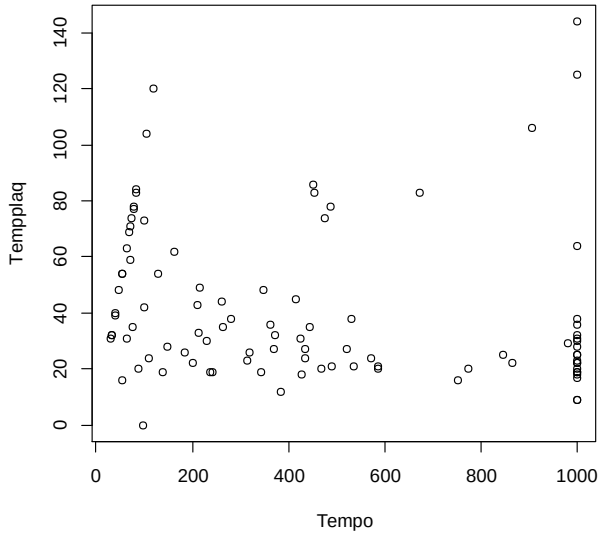

Figura 9a: Gráfico de dispersão do tempo de recuperação de plaquetas dos pacientes ao longo do tempo de sobrevivência.

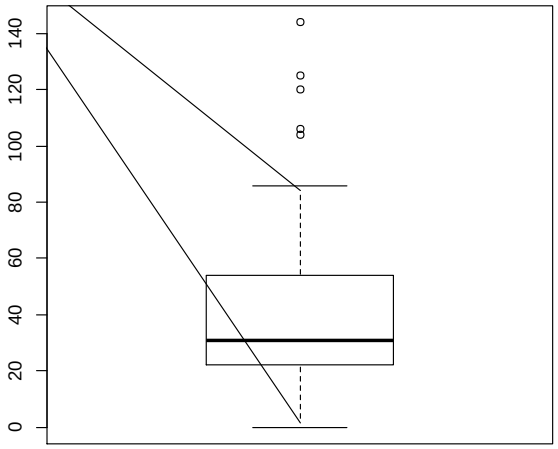

Figura 9b: Boxplot do tempo de recuperação de plaquetas dos pacientes.

Através da Figura 9a e da Figura 9b, observa-se que os tempos de recuperação de plaquetas se concentram no intervalo (20,60) e a presença de indivíduos censurados no Tempo=1000. Pelo gráfico de dispersão também nota-se que não existe um indicativo claro de correlação entre os tempos de sobrevivência e os tempos de recuperação de plaquetas nos pacientes. Através do boxplot também se verificou a presença de alguns *outiliers*.

- X_{i5} : Desenvolvimento da Doença Enxerto Aguda

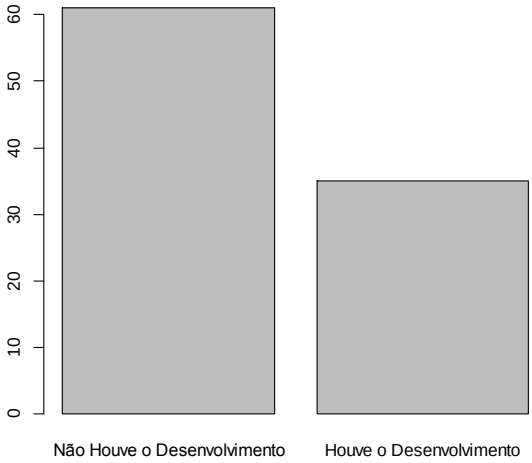

Figura 10a: Gráfico da frequência dos pacientes considerando o desenvolvimento da doença enxerto aguda.

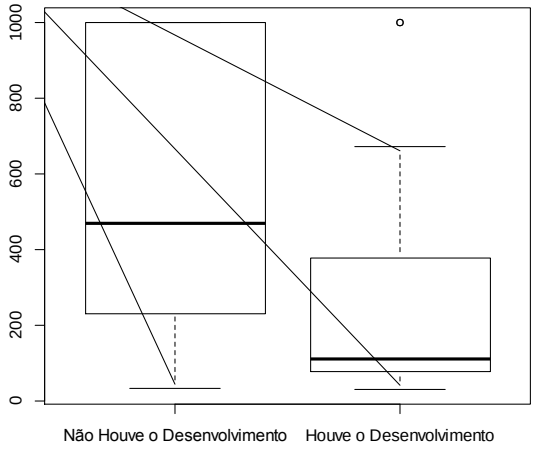

Figura 10b: Boxplot dos tempos de sobrevivência de cada grupo considerando o desenvolvimento da doença enxerto aguda.

A Figura 10a mostra que existe uma predominância de indivíduos que não tiveram o desenvolvimento da doença enxerto aguda. A Figura 10b evidencia que a covariável influencia nos tempos de sobrevivência, dado que os boxplot's possuem comportamentos diferentes.

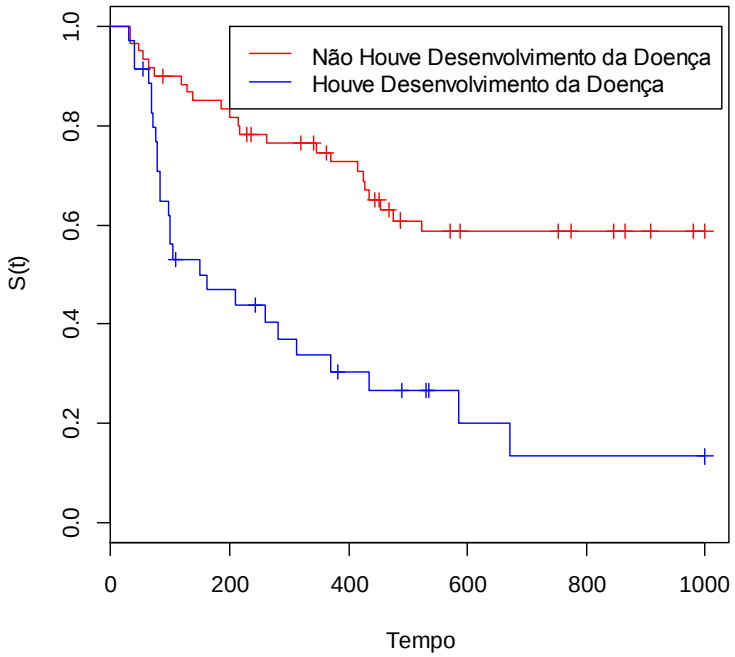

Figura 11: Gráfico da função de sobrevivência dos pacientes, considerando a variável X_5, estimada pelo método de Kaplan-Meier.

A Figura 11 indica que as curvas são diferentes, de forma que os pacientes que desenvolveram a doença têm menor probabilidade de sobreviver. Para confirmar estas

afirmações, novamente realizou-se o teste de Log-Rank. Nota-se um cruzamento das curvas, o que poderia significar a não proporcionalidade dos riscos. Apesar disso, utilizou-se o teste de Log-Rank já que o cruzamento é pequeno e ocorre apenas bem no início do estudo.

Com p-valor < 0.0001, rejeitou-se a hipótese nula e assim há evidências para considerar que as curvas de sobrevivência dos dois grupos são realmente diferentes, o que confirma a conclusão preliminar feita através da Figura 10b. Calculou-se o risco relativo e concluiu-se que os pacientes que não desenvolveram a doença têm 3.18 mais chances de sobreviver que os outros pacientes.

- X_{i6}: Tempo de Desenvolvimento da Doença Enxerto Aguda

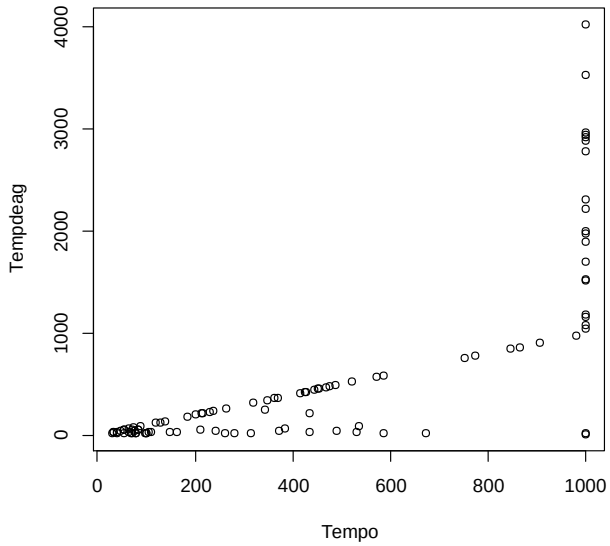

Figura 12a: Gráfico de dispersão do tempo até o desenvolvimento da doença enxerto aguda dos pacientes ao longo do tempo de sobrevivência.

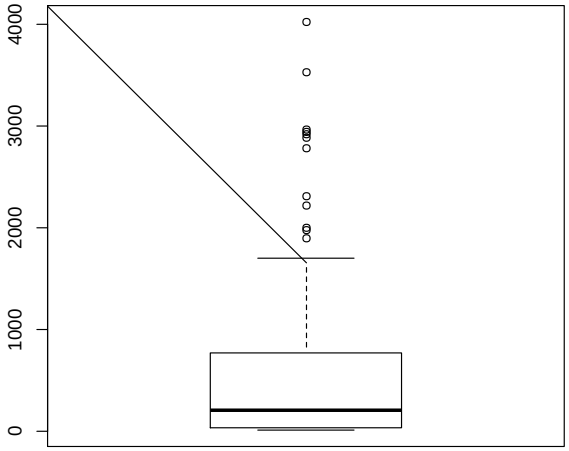

Figura 12b: Boxplot do tempo até o desenvolvimento da doença enxerto aguda dos pacientes.

Através da Figura 12a, observa-se que existe uma clara correlação positiva entre os tempos de desenvolvimento da doença e os tempos de sobrevivência. Também nota-se que há uma concentração de valores baixos no início do estudo e que há vários indivíduos censurados no Tempo=1000. A Figura 12b indica que há uma concentração de valores abaixo de 1000 e que há *outliers*.

- X_{i7}: Desenvolvimento da Doença Enxerto Crônica

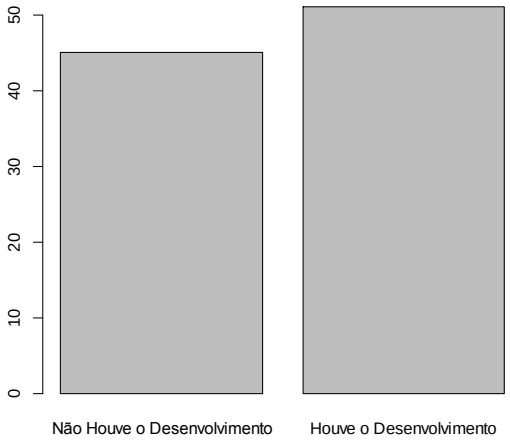

Figura 13a: Gráfico da frequência dos pacientes considerando o desenvolvimento da doença enxerto crônica.

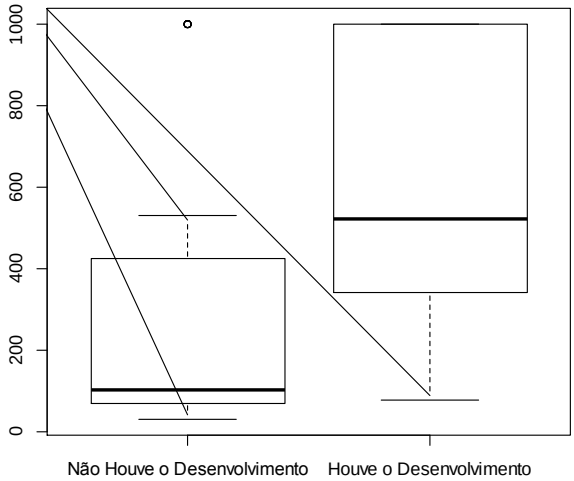

Figura 13b: Boxplot dos tempos de sobrevivência de cada grupo considerando o desenvolvimento da doença enxerto crônica.

A Figura 13a mostra que não há muita diferença entre o número de pacientes que desenvolveram a doença de enxerto crônica e que não desenvolveram. Ao comparar os boxplot's da Figura 13b, nota-se uma diferença no comportamento dos grupos, indicando assim que pode haver uma influência dessa covariável nos tempos de sobrevivência.

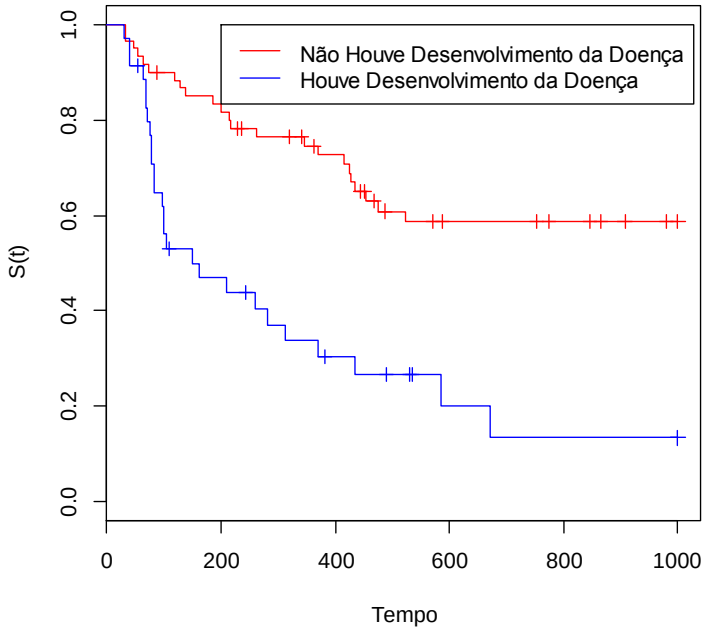

Figura 14: Gráfico da função de sobrevivência dos pacientes, considerando o desenvolvimento da doença enxerto crônica, estimada pelo método de Kaplan-Meier.

Observa-se pela Figura 14 que a curva de sobrevivência para os dois grupos são diferentes e que os pacientes que não desenvolveram a doença têm mais probabilidade de sobreviver. Com p-valor < 0.0001, pelo teste de Log-Rank, rejeita-se a hipótese nula, tendo assim evidências para acreditar que as curvas dos grupos realmente são diferentes. Assim como foi realizado considerando a covariável X_5, existe um cruzamento inicial e pequeno das curvas, que por hora não será considerado para descartar o indício de riscos proporcionais e o uso do teste de Log-Rank. A chance de sobreviver dos pacientes que não desenvolveram a doença é aproximadamente 3,66 vezes maior do que daqueles que desenvolveram.

- X_{i8}: Tempo de Desenvolvimento da Doença Enxerto Crônica

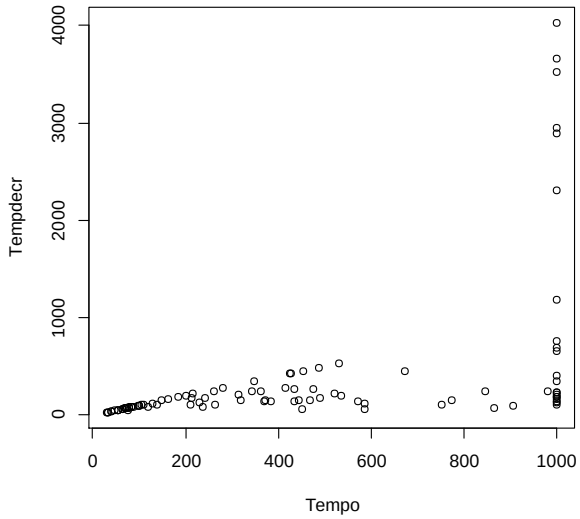

Figura 15a: Gráfico de dispersão do tempo até o desenvolvimento da doença enxerto crônica.

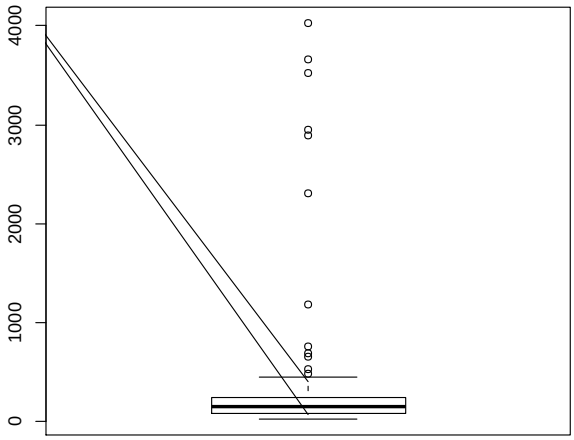

Figura 15b: Boxplot do tempo até o desenvolvimento da doença enxerto crônica..

Através da Figura 15a, observa-se que o tempo para o desenvolvimento da doença de enxerto crônica aumenta levemente ao longo do tempo, indicando assim que pode haver uma correlação entre os tempos de sobrevivência e o tempo de desenvolvimento da doença enxerto crônica. Existe também uma leve concentração de valores baixos no início do estudo e observam-se indivíduos censurados no Tempo=1000. A Figura 15b indica que há uma predominância de valores baixos dos tempos de desenvolvimento da doença e que existem alguns *outliers* na amostra.

- Fase

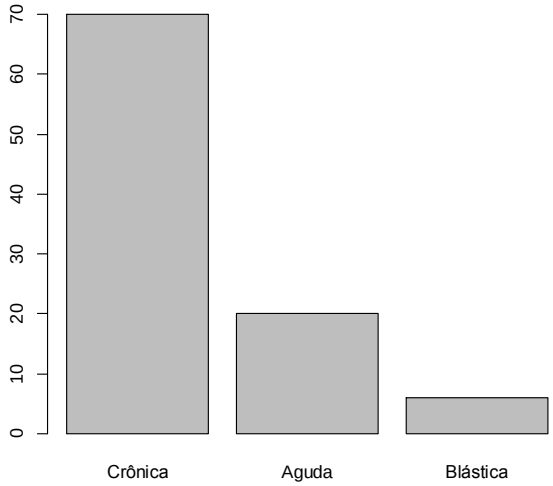

Figura 16a: Gráfico da frequência dos pacientes dada a fase em que eles estavam quando foi feito o transplante.

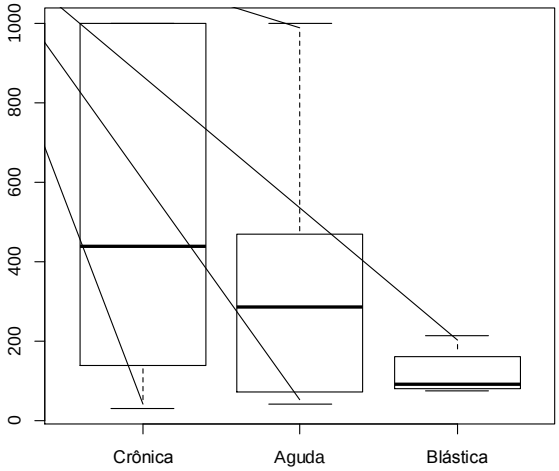

Figura 16b: Boxplot dos tempos de sobrevivência de cada grupo dada a fase em que eles estavam quando foi feito o transplante.

A Figura 16a mostra que há uma predominância de pacientes que foram diagnosticados na fase crônica. Ao analisar a Figura 16b, nota-se uma grande diferença entre os boxplot's de cada grupo, o que indica que a variável influencia os tempos de sobrevivência dos indivíduos.

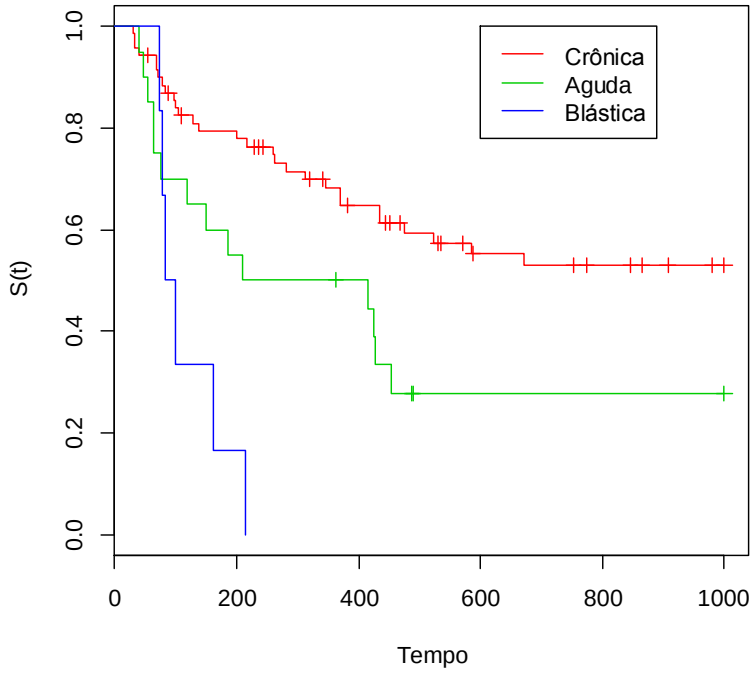

Figura 17: Gráfico da função de sobrevivência dos pacientes, considerando a covariável X_9, estimada pelo método de Kaplan-Meier.

O gráfico indica que as curvas de sobrevivência para os três grupos diferem e que os pacientes que se encontravam na fase blástica tinham menor probabilidade de sobreviver, seguido pelos pacientes que se encontravam na fase aguda e depois crônica. Confirmou-se essa suposição com o teste de Log-Rank, obtendo-se um p-valor < 0,0001.

Como existem três grupos dentro dessa variável, realizou-se o teste 2 a 2 para verificar quais dos grupos realmente se diferem. Primeiramente, testou-se a hipótese de que os grupos de pacientes que se encontravam nas fases crônica e aguda têm a mesma curva de sobrevivência. O teste de Log-Rank foi realizado e, com p-valor < 0.0001, rejeitou-se a hipótese, considerando assim que as curvas de sobrevivência dos dois grupos não são iguais.

Ao comparar as curvas dos pacientes que se encontravam nas fases crônica e blástica, a partir do teste de Log-Rank, encontrou-se um p-valor aproximadamente igual a 0.0007, rejeitando assim a hipótese e considerando que as curvas dos pacientes que se encontravam nessas duas fases não são iguais.

Por último, compararam-se as curvas dos pacientes que se encontravam nas fases aguda e blástica e, com o teste de Log-Rank, encontrou-se um p-valor de aproximadamente 0.004, rejeitando-se assim a hipótese e considerando que as curvas dos dois grupos também não são iguais.

3.2 Ajuste do Modelo

O estudo foi feito a partir de uma amostra de 96 pacientes que realizaram transplante de medula óssea no Cemo-Inca. O foco foi analisar os tempos de sobrevivência desses pacientes, isto é, o tempo que cada um levou desde o transplante de medula, até o óbito.

Em pesquisas, quando o interesse é explicar o comportamento de alguma variável específica, utilizam-se modelos de regressão. Parte-se das respostas da variável em questão de uma amostra e coletam-se dados referentes a outras variáveis que possam de alguma maneira influenciar nessa variável principal.

Nesse estudo conta-se com a presença de pacientes que não concluíram o evento de interesse, que é o óbito. Para modelar então o conjunto de dados, utilizou-se a metodologia de regressão de análise de sobrevivência, que possibilita a ocorrência de indivíduos censurados.

A partir da análise descritiva das variáveis estudadas, verificou-se que várias variáveis parecem influenciar nos tempos de sobrevivência dos pacientes. Também foi observado que existe um forte indicativo de riscos proporcionais, o que leva a crer que ajustar os dados a partir do modelo de Cox é uma boa alternativa.

Para avaliar se os dados realmente poderiam ser ajustados pelo modelo de Cox, verificou-se a suposição de riscos proporcionais para cada variável. Para tal, foram ajustados modelos considerando cada covariável para explicar os tempos de sobrevivência. A seguir, as conclusões para cada uma dessas variáveis individualmente:

Tabela 2: Estatísticas a respeito do Risco Proporcional das Variáveis.

Variável	ρ	p-valor
X_1	-0.197	0.169
X_2	0.117	0.451
X_3	-0.00271	0.984
X_4	-0.192	0.364
X_5	-0.0219	0.879
X_6	0.536	0.00569
X_7	0.472	0.00224
X_8	0.744	<0,0001
X_{92}	-0.0203	0.888
X_{93}	0.0811	0.557
X_9 Global	-	0.806

Sendo assim, de acordo com as estatísticas da Tabela 2, concluiu-se que as variáveis que possuem riscos proporcionais são: $X_1, X_2, X_3, X_4, X_5, X_{92}, X_{93}, X_9 Global$.

Depois dessa verificação, para a escolha das variáveis que deveriam estar no modelo final, os dados foram ajustados ao modelo de Cox considerando todas as variáveis estudadas. As estatísticas desse modelo inicial estão a seguir:

Tabela 3: Estatísticas do modelo inicial que contém todas as variáveis em estudo.

Variável	Parâmetro	Estimativa	Z	p-valor (z)	ρ	p-valor (ρ)
X_1	β_1	-0.046472	-0.138	0.88986	-0.1083	0.43912
X_2	β_2	-0.035304	-1.975	0.04831	-0.1783	0.11397
X_3	β_3	-1.636778	-2.630	0.00854	-0.0138	0.91272
X_4	β_4	0.004289	0.599	0.54906	0.1204	0.42411
X_5	β_5	-1.404991	-3.118	0.00182	-0.258	0.05578
X_6	β_6	-0.006003	-4.333	1.47E-05	-0.0835	0.64016
X_7	β_7	-2.295576	-4.473	7.71E-06	0.2311	0.05969
X_8	β_8	-0.008468	-4.019	5.85E-05	0.4932	0.00174
X_{92}	β_{92}	0.951221	2.486	0.01293	-0.0357	0.78879
X_{93}	β_{93}	0.455199	0.818	0.41329	0.0646	0.60046
Global	-	-	-	-	-	0.00641

Observa-se pelas estatísticas da Tabela 3 que este modelo não possui riscos proporcionais e que nem todas as variáveis são significativas.

Para chegar ao modelo final, retirou-se uma variável por vez, sempre aquela com menor p-valor para a suposição de riscos proporcionais. Ao mesmo tempo, controlou-se na seleção a significância das variáveis no modelo. Ao fazer isso, observou-se que cada retirada de variável modificou as estatísticas das outras variáveis.

Após analisar diversos modelos, com diversas combinações de variáveis, o modelo final, que melhor representa os dados e que sustenta todas as hipóteses necessárias para que o modelo seja validado é explicado pelas variáveis X_3, X_6 e X_9:

Tabela 4: Estatísticas do modelo considerando as variáveis X_3, X_6 e X_9:

Variável	Parâmetro	Estimativa ($\widehat{\beta}$)	$Exp\{\hat{\beta}\}$	Z	p-valor (z)	95% I.C.($Exp\{\hat{\beta}\}$)
X_3	β_1	-2.7591325	0.0634	-5.462	<0,0001	(0.0235 , 0.1705)
X_6	β_2	-0.0033899	0.9966	-4.251	<0,0001	(0.9951 , 0.9982)
X_{92}	β_3	0.7763370	2.1735	2.248	0.0246	(1.1045 ,4.2770)
X_{93}	β_4	1.0087191	2.7421	2.043	0.0410	(1.0419 , 7.2164)

Estatística	Valor	Estatística	Valor	p-valor (g.l. = 4)
R^2	0,688	Razão da Verossimilhança	111.7	0

Tabela 5: Estatísticas a respeito do Risco Proporcional das variáveis selecionadas

Variável	Parâmetro	ρ	p-valor (ρ)
X_3	β_1	-0.0213	0.8752
X_6	β_2	0.3733	0.0621
X_{92}	β_3	0.0276	0.8441
X_{93}	β_4	0.1041	0.4367
Global	-	-	0.3617

A partir das estatísticas acima, assume-se que o modelo final possui riscos proporcionais, inclusive cada covariável individualmente, e que todas as variáveis são significantes.

A função risco do modelo é dada por:

$$h_t = h_0 \exp\{-2.7591325\ X_3 - 0.0033899\ X_6 + 0.7763370\ X_{92} + 1.0087191\ X_{93}\}.$$

A partir dessa taxa de risco, conclui-se que:

- Se não houver recuperação de plaquetas, o risco de falha do indivíduo é $\exp\{-2{,}75913\}^{-1} \cong 15{,}8$ vezes maior que o risco dos indivíduos que tiveram recuperação de plaquetas;

- Quanto maior o tempo de desenvolvimento da doença de enxerto aguda, menor o risco de o indivíduo falhar; $\exp\{-2{,}75913\}^{-1} \cong 15{,}8$

- Considerando a fase crônica como a fase controle, tem-se que:

    - Se a fase em que o paciente se encontrava quando foi diagnosticado com a doença era a fase aguda, este paciente tem taxa de risco de falha $\exp\{0,77634\} \cong 1,5842$ vezes maior que o paciente que fez o transplante na fase crônica;

    - Se a fase em que o paciente se encontrava quando foi diagnosticado com a doença era a fase blástica, este paciente tem taxa de risco de falha $\exp\{1,00872\} \cong 2,74$ vezes maior que o paciente que fez o transplante na fase crônica.

3.3 Análise de Diagnóstico

Após ajustar o modelo, é importante verificar a qualidade desse ajuste. Para tal, fez-se uso das metodologias de Análise de Resíduos e de Análise de Influência Global.

3.3.1 Análise de Resíduos

Utilizaram-se os resíduos Martingal e Deviance para verificar se os dados forem bem ajustados pelo modelo e se existem observações influentes. Os gráficos dos resíduos estão a seguir:

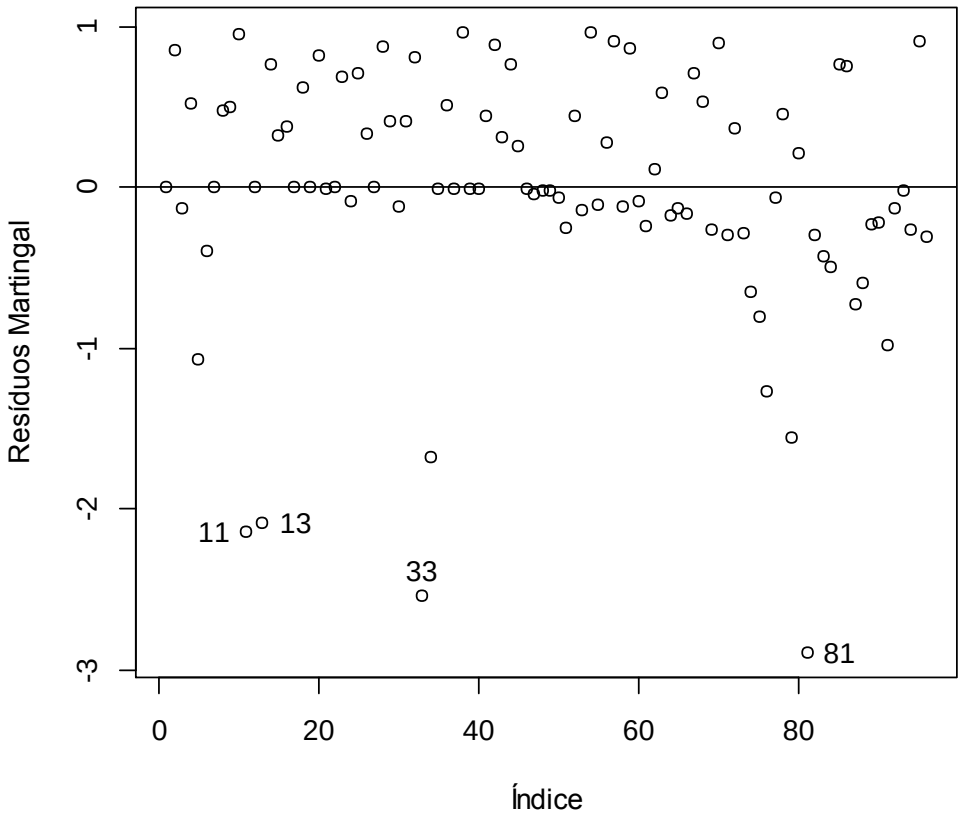

Figura 18: Gráfico dos resíduos Martingal.

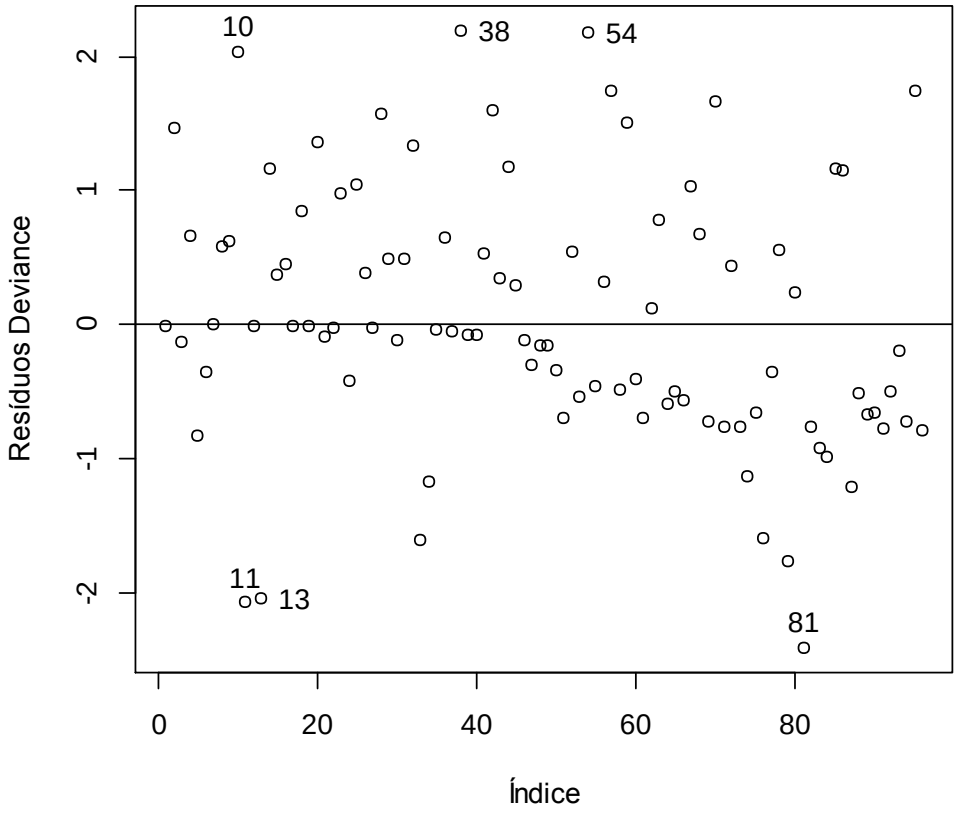

Figura 19: Gráfico dos Resíduos Deviance

Analisando o gráfico de resíduos Martingal, observa-se um comportamento mais aleatório dos resíduos acima do valor zero, o que é esperado se o modelo for bem ajustado.

O comportamento não aleatório dos resíduos abaixo do valor zero exige uma maior atenção aos dados referentes a esses resíduos. Ao analisar os dados individualmente, observou-se que os dados que estão próximos ao valor zero e abaixo são, em sua maioria, observações censuradas que possuem o mesmo tempo de sobrevivência. Os indivíduos 11, 13, 33 e 81 são os que mais se distanciaram dos outros resíduos, o que pode ser um indício de que estes indivíduos são influentes.

Agora, ao analisar o gráfico de resíduos Deviance, existe um comportamento bem aleatório dos resíduos em torno do valor zero. Verificou-se que as observações 10, 11, 13, 38, 54 e 81 são possíveis observações influentes por se distanciarem das demais.

3.3.2 Análise de Influência Global

Outra metodologia importante para verificar se há observações influentes nos dados é a de deleção de casos. Utilizaram-se os gráficos do Afastamento da Verossimilhança e da Distância de Cook para tal, que estão a seguir:

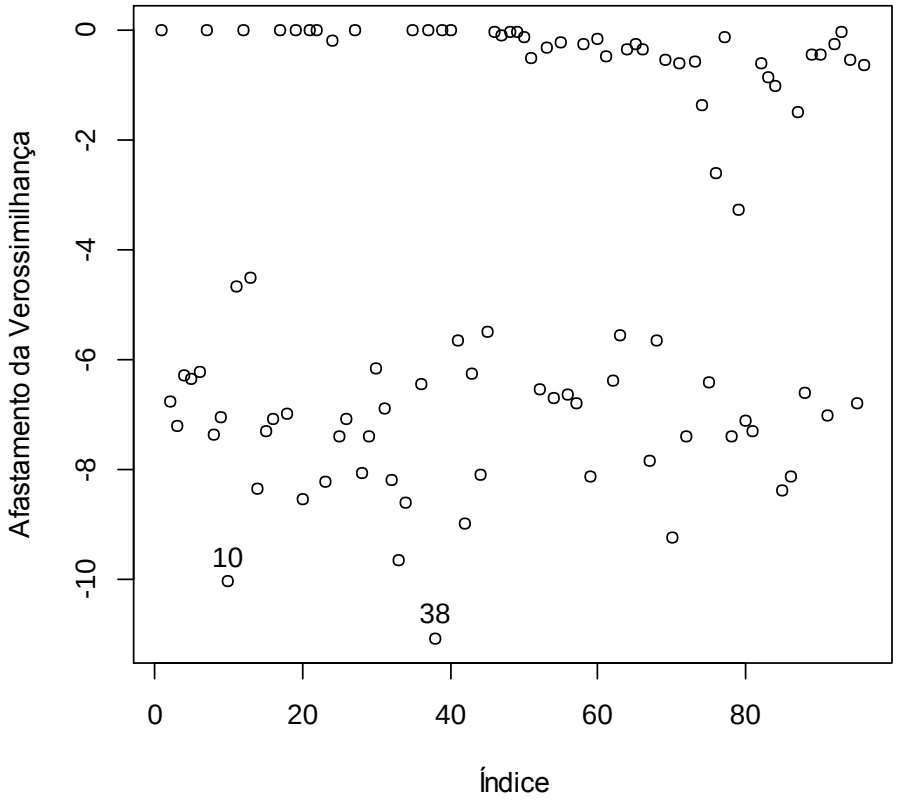

Figura 20: Gráfico de dispersão das medidas de Afastamento da Verossimilhança.

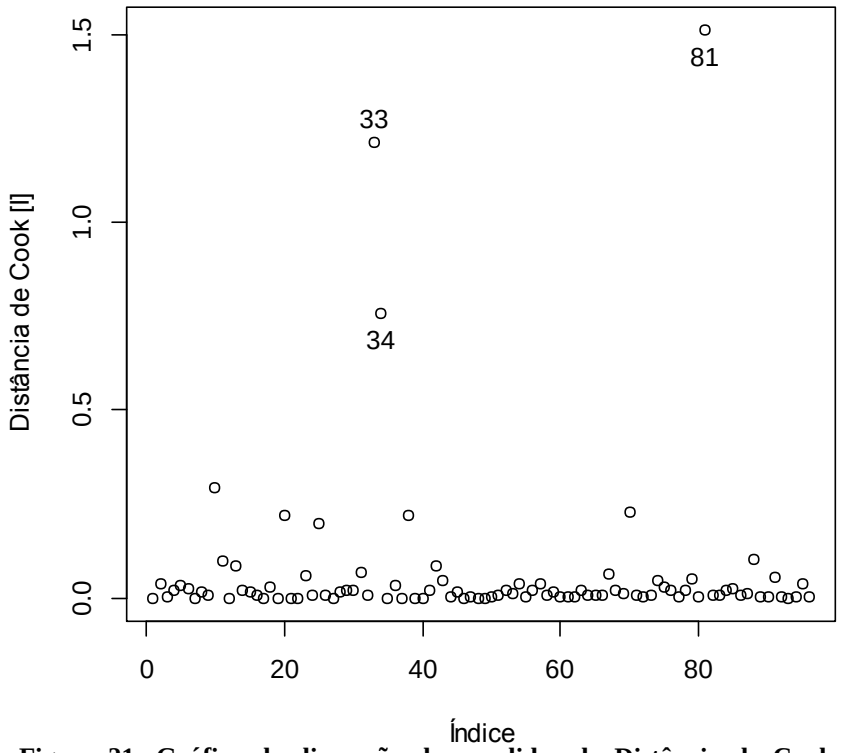

Figura 21: Gráfico de dispersão das medidas da Distância de Cook calculadas a partir da matriz Hessiana.

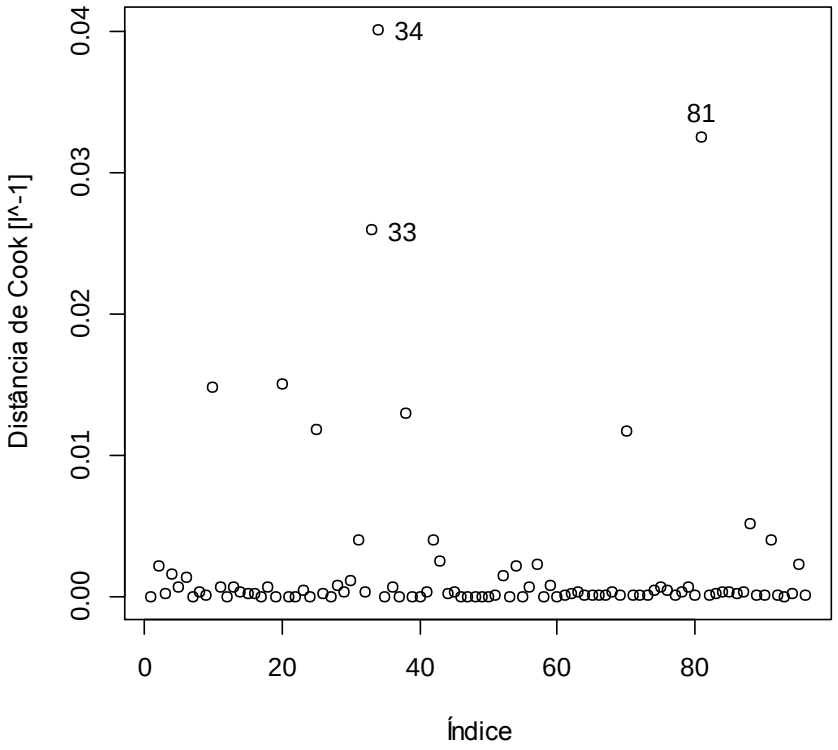

Figura 22: Gráfico de dispersão das medidas da Distância de Cook calculadas a partir do inverso da matriz Hessiana.

A partir da Figura 20, do gráfico do Afastamento da Verossimilhança, foi possível verificar que as observações 10 e 38 se distanciam das demais.

A partir Figura 21 e da Figura 22, do gráfico da Distância de Cook, tanto o gráfico calculado a partir da matriz hessiana quanto o calculado pela inversa da hessiana, as observações que mais se afastaram das demais foram as 33, 34 e 81.

Dessa maneira, esses indivíduos requerem uma maior atenção no ajuste do modelo, devendo haver uma avaliação sobre eles para julgar se esses devem ser mantidos no estudo ou não.

3.4 Análise das Observações Influentes

Após realizar a análise de diagnóstico no modelo de Cox aplicado aos dados dos pacientes através da metodologia de Análise de Resíduos e de Influência Global, verificou-se que as observações #33, #34 e #81 podem ser influentes. A seguir, uma tabela com as informações de cada paciente:

Tabela 6: Resposta das variáveis explicativas dos pacientes 33, 34 e 81.

Indivíduo (i)	δ_i	T_i	X_{i1}	X_{i2}	X_{i3}	X_{i4}	X_{i5}	X_{i6}	X_{i7}	X_{i8}	X_{i9}
33	1	120	1	38	0	120	0	120	1	87	2
34	1	83	1	33	0	83	1	55	0	83	3
81	0	489	2	43	1	21	1	44	1	181	2

Ao analisar a Tabela 6, à primeira vista, pode-se constatar que o tempo para a falha do primeiro indivíduo (33) é relativamente alto dado à falta da recuperação de plaquetas e o fato de que ele fez o transplante estando na fase aguda. O segundo indivíduo (34) também teve um tempo de falha relativamente grande considerado o tempo até o desenvolvimento da doença enxerto aguda, a não recuperação de plaquetas e que estava na fase blástica na época do transplante. Já o último indivíduo (81) é censurado, o que também pode indicar alguma inconformidade, dado que ele se encontrava na fase aguda da doença quando fez o transplante.

Para melhor averiguar se esses pacientes estão influenciando o modelo, estimaram-se novos parâmetros a partir de novas amostras. Essas novas amostras são na verdade sub-amostras da amostra original, que foram selecionadas retirando-se cada observação dita como influente e todas as combinações possíveis entre elas. Dessa maneira, os parâmetros dos modelos estimados a partir dessas novas amostras mudam e, consequentemente, os seus p-valores e outras estatísticas referentes a eles também. Para essa análise também se utilizou o valor da mudança relativa do modelo, definida como RC.

A tabela a seguir possui a mudança relativa de cada ajuste para cada sub-amostra, seguida pela estimativa do parâmetro, p-valor do parâmetro, p-valor de ρ e mais algumas estatísticas dos novos ajustes:

Tabela 7: [Mudança relativa], estimativa do parâmetro, (p-valor do parâmetro), {p-valor de ρ (riscos proporcionais)} e mais algumas estatísticas de cada modelo de cada sub-amostra.

Sub-Amostra	β_1 [-]	β_2 [-]	β_3 [-]	β_4 [-]	P-valor R.P. *	Verossimilhança
I - completo	-2,75913 (<0,0001) {0,8752}	-0,00339 (<0,0001) {0,0621}	0,77634 (0,0246) {0,8441}	1,00872 (0,041) {0,4367}	0,3617	-147,5498
I - {33}	[-0,09794] -3,02936 (<0,0001) {0,9146}	[0,00322] -0,00338 (<0,0001) {0,0856}	[-0,36240] 1,05768 (0,00321) {0,6696}	[0,08984] 0,91809 (0,06357) {0,5238}	0,4764	-142,729
I - {34}	[-0,04536] -2,88428 (<0,0001) {0,8232}	[0,00581] -0,00337 (<0,0001) {0,0679}	[0,00903] 0,76933 (0,02606) {0,837}	[-0,38858] 1,40068 (0,00741) {0,4839}	0,3959	-143,2558
I - {81}	[0,04909] -2,62369 (<0,0001) {0,9408}	[-0,16756] -0,00396 (<0,0001) {0,0781}	[-0,49820] 1,16311 (0,00223) {0,4491}	[-0,12141] 1,13119 (0,02424) {0,3423}	0.1989	-143,9026
I - {33,34}	[-0,15042] -3,17415 (<0,0001) {0,8502}	[0,00858] -0,00336 (<0,0001) {0,0937}	[-0,38218] 1,07304 (0,00299) {0,6689}	[-0,37068] 1,38263 (0,00846) {0,5771}	0,5179	-138,2881
I - {33,81}	[-0,09414] -3,01889 (<0,0001) {0,8239}	[-0,22192] -0,00414 (<0,0001) {0,0924}	[-1,01639] 1,5654 (<0,0001) {0,9075}	[-0,02706] 1,03601 (0,0399) {0,4735}	0,3794	-138,0499
I - {34,81}	[0,00188] -2,75394 (<0,0001) {0,8898}	[-0,15803] -0,00393 (<0,0001) {0,0856}	[-0,47826] 1,14763 (0,00255) {0,4418}	[-0,48621] 1,49917 (0,00471) {0,3949}	0,2226	-139,7021
I - {33,34,81}	[-0,15203] -3,17861 (<0,0001) {0,754}	[-0,21847] -0,00413 (<0,0001) {0,102}	[-1,04190] 1,5852 (<0,0001) {0,902}	[-0,49578] 1,50882 (0,00482) {0,512}	0,419	-133,6377

P-Valor R.P. *: Se refere ao p-valor de ρ, utilizado para verificar se os riscos são proporcionais.

A partir da Tabela 7 admite-se que a significância dos parâmetros se manteve em todos os modelos criados a partir das sub-amostras, assim como a presença de riscos proporcionais. Também se observou que os valores das mudanças relativas foram pequenos.

Apesar disso, é possível notar que a estimativa $\hat{\beta}_4$ variou quando a observação #33 foi retirada e, por mais que a mudança tenha sido pequena, sua interpretação mudou. Já a interpretação da estimativa $\hat{\beta}_3$ oscilou bastante de acordo com cada sub-amostra considerada.

Sendo assim, conclui-se que o modelo é sensível a essas observações, especialmente a #33 e a #34. Seria prudente ter uma conversa com o pesquisador para averiguar se essas observações devem ou não ser mantidas no modelo.

4 CONCLUSÃO

Neste trabalho foi proposto analisar os tempos de sobrevivência de uma coorte de 96 pacientes que foram submetidos a um transplante de medula óssea para o tratamento de leucemia mielóide crônica.

Nove variáveis foram consideradas para explicar os tempos de sobrevivência dos pacientes. Visto que foi verificada a presença de riscos proporcionais na maioria dessas variáveis, os dados foram ajustados ao modelo de Cox.

Entre as nove variáveis observadas, apenas três entraram no modelo, sendo essas a recuperação de plaquetas, o tempo de desenvolvimento da doença de enxerto aguda e a fase em que o paciente se encontrava quando o transplante foi feito, X_3, X_6 *e* X_9 respectivamente. Essas variáveis foram escolhidas considerando tanto a presença de riscos proporcionais, como a significância dessas.

O modelo foi ajustado e a partir dele verificou-se que o indivíduo que tiver recuperação de plaquetas tem taxa de risco aproximadamente 0,06 vezes menor que o risco dos indivíduos que não tiveram essa recuperação. Observou-se também que quanto maior o tempo até o desenvolvimento da doença de enxerto aguda, menor o risco de o indivíduo falhar. Se o paciente se encontrava na fase aguda quando foi feito o transplante, ele tem taxa de risco aproximadamente 1,6 vezes maior que o paciente que fez o transplante na fase crônica e que se o paciente se encontrava na fase blástica, o risco é 2,7 vezes maior que o risco do paciente que se encontrava na fase crônica.

Após esse ajuste, foi feito uma análise de diagnóstico para avaliar a qualidade do ajuste, averiguar se havia observações influentes no modelo. As metodologias de Análise de Resíduos e Influência Global foram utilizadas. Através delas, verificou-se que as observações #33, #34 e #81 poderiam ser influentes.

Sendo assim, foram criadas sete sub-amostras e em cada uma delas foi retirada uma das observações ditas influentes e suas combinações. Utilizou-se o valor da mudança relativa para avaliar se a mudança na estimativa foi grande ou não. Os p-valores também foram analisados, assim como os valores das verossimilhanças de cada modelo.

Depois dessa análise, verificou-se que a significância dos parâmetros se manteve em todos os modelos, assim como a presença de riscos proporcionais. Além disso, os valores das mudanças relativas foram pequenos.

Um fato que exige uma maior atenção é que a estimativa$\hat{}$ teve uma mudança na interpretação quando a observação #33 foi retirada. A interpretação da estimativa$\hat{}$ oscilou mais considerando os modelos criados a partir das sub-amostras. Ambos os casos podem ser consequência das inconformidades dos tempos de sobrevivência dos indivíduos considerando a fase em que estavam na data do transplante. Dessa maneira, seria adequado ter uma conversa com o pesquisador para avaliar se seria melhor a retirada de alguma dessas observações consideradas influentes.

5 ANEXOS

A seguir a programação utilizada no software R:

```
dados<-read.csv(file.choose())

library(survival)

tempo<-dados$tempo

censura<-dados$status

sexo<-dados$sexo

idade<-dados$idade

plaq<-dados$plaq

plot(plaq)

boxplot(tempo~plaq)

tempplaq<-dados$tempplaq

deag<-dados$deag

tempdeag<-dados$tempdeag

decr<-dados$decr

tempdecr<-dados$tempdecr

fase<-dados$fase


#Análise Descritiva das Variáveis

sexo<-factor(sexo, levels=c(1,2), labels=c("Masculino", "Feminino"))

plaq<-factor(plaq, levels=c(0,1), labels=c('Não Houve Recuperação', 'Houve Recuperação'))
```

```
deag<-factor(deag, levels=c(0,1), labels=c('Não Houve o Desenvolvimento','Houve o Desenvolvimento'))

decr<-factor(decr, levels=c(0,1), labels=c('Não Houve o Desenvolvimento', 'Houve o Desenvolvimento'))

fase<-factor(fase, levels=c(1,2,3), labels=c('Crônica', 'Aguda', 'Blástica'))


#Sexo

boxplot(tempo~sexo)

barplot(table(sexo))


#Idade

boxplot(idade)

plot(tempo,idade,xlab='Tempo',ylab='Idade')


#Plaq

boxplot(tempo~plaq)

barplot(table(plaq))


#Tempplaq

boxplot(tempplaq)

plot(tempo,tempplaq,xlab='Tempo',ylab='Tempplaq')


#Deag
```

```
boxplot(tempo~deag)

barplot(table(deag))


#Tempdeag

boxplot(tempdeag)

plot(tempo,tempdeag,xlab='Tempo',ylab='Tempdeag')


#Decr

boxplot(tempo~decr)

barplot(table(decr))


#Tempdecr

boxplot(tempdecr)

plot(tempo,tempdecr,xlab='Tempo',ylab='Tempdecr')


#Fase

boxplot(tempo~fase)

barplot(table(fase))


#Análise de Sobrevivência

 summary(dados)

 tempo<-dados$os

 length(tempo)
```

```
 hist(tempo,xlab="Tempo de Sobrevivência", ylab="Frequência de Pacientes", main="")
 title(main="Frequência de Pacientes x Tempo de Sobrevivência")

KM<-survfit(Surv(tempo,censura)~1, conf.int=F)
summary(KM)
plot(KM,conf.int=T, xlab="Tempo", ylab="S(t)")
title(main="Curva de Sobrevivência")
plot(KM, conf.int=F, fun="cumhaz", xlab="Tempo", ylab="H(t)")
title(main="Função Risco Acumulado")

#SEXO
KMs<-survfit(Surv(tempo,censura)~sexo, conf.int=F)
plot(KMs, conf.int=F, xlab="Tempo", ylab="S(t)", lty=c(1,1),col=c(4,2))
legend(500,1, lty=c(1,1),col=c(4,2), c('Masculino', 'Feminino'))
survdiff(Surv(tempo, censura)~sexo, rho=0)
survdiff(Surv(tempo, censura)~sexo, rho=1)

#RECUPERAÇÃO DE PLAQUETAS
KMp<-survfit(Surv(tempo,censura)~plaq, conf.int=F)
plot(KMp, conf.int=F, xlab="Tempo", ylab="S(t)", lty=c(1,1),col=c(2,4))
legend(220,0.4, lty=c(1,1), col=c(2,4), c('Não Houve Recuperação de Plaquetas', 'Houve
Recuperação de Plaquetas'))
```

```
survdiff(Surv(tempo, censura)~plaq, rho=0)

survdiff(Surv(tempo, censura)~plaq, rho=1)



difere<-survdiff(Surv(tempo,censura)~plaq, rho=0)

names(difere)   #mostra todos os "names" relacionados a função difere#

difere$obs    #obs das duas linhas#

difere$obs[1] #observações linha 1#

difere$obs[2] #Obs linha 2#

RR<-(difere$obs[1]/difere$exp[1])/(difere$obs[2]/difere$exp[2])

RR #para sair o resultado, no caso 11,29#



#DOENÇA ENXERTO AGUDA

KMdeag<-survfit(Surv(tempo,censura)~deag, conf.int=F)

plot(KMdeag, conf.int=F, xlab="Tempo", ylab="S(t)", lty=c(1,1),col=c(2,4))

legend(200,1, lty=c(1,1),col=c(2,4), c('Não Houve Desenvolvimento da Doença', 'Houve Desenvolvimento da Doença'))


survdiff(Surv(tempo, censura)~deag, rho=0)

survdiff(Surv(tempo, censura)~deag, rho=1)
```

```
difere<-survdiff(Surv(tempo,censura)~deag, rho=0)

names(difere)   #mostra todos os "names" relacionados a função difere#

difere$obs    #obs das duas linhas#

difere$obs[1] #observações linha 1#

difere$obs[2] #Obs linha 2#

RR<-(difere$obs[1]/difere$exp[1])/(difere$obs[2]/difere$exp[2])

RR



#DOENÇA ENXERTO CRÔNICA

KMdecr<-survfit(Surv(tempo,censura)~deag, conf.int=F)

plot(KMdeag, conf.int=F, xlab="Tempo", ylab="S(t)", lty=c(1,1),col=c(2,4))

legend(200,1,  lty=c(1,1),col=c(2,4),  c('Não Houve Desenvolvimento da Doença', 'Houve Desenvolvimento da Doença'))


survdiff(Surv(tempo, censura)~decr, rho=0)

survdiff(Surv(tempo, censura)~decr, rho=1)


difere<-survdiff(Surv(tempo,censura)~decr, rho=0)

names(difere)   #mostra todos os "names" relacionados a função difere#

difere$obs    #obs das duas linhas#

difere$obs[1] #observações linha 1#

difere$obs[2] #Obs linha 2#
```

```
RR<-(difere$obs[1]/difere$exp[1])/(difere$obs[2]/difere$exp[2])

RR


#FASE

KMf<-survfit(Surv(tempo,censura)~fase, conf.int=F)

plot(KMf, conf.int=F, xlab="Tempo", ylab="S(t)", lty=c(1,1,1),col=c(2,3,4))

legend(600,1, lty=c(1,1,1),col=c(2,3,4), c('Crônica', 'Aguda', 'Blástica'))


survdiff(Surv(tempo, censura)~decr, rho=0)

survdiff(Surv(tempo, censura)~decr, rho=1)


difere<-survdiff(Surv(tempo,censura)~decr, rho=0)

names(difere)   #mostra todos os "names" relacionados a função difere#

difere$obs    #obs das duas linhas#

difere$obs[1] #observações linha 1#

difere$obs[2] #Obs linha 2#

RR<-(difere$obs[1]/difere$exp[1])/(difere$obs[2]/difere$exp[2])

RR


fase<-factor(dados$fase)

length(fase)

dadosfase<-dados[order(dados$fase),]

summary(as.factor(dados$fase))
```

```
survdiff(Surv(c(tempo[1:70],tempo[71:90]),c(censura[1:70],censura[71:90]))~c(fase[1:70],fase[71:90]), rho=0)

survdiff(Surv(c(tempo[1:70],tempo[91:96]),c(censura[1:70],censura[91:96]))~c(fase[1:70],fase[91:96]), rho=0)

survdiff(Surv(c(tempo[71:90],tempo[91:96]),c(censura[71:90],censura[91:96]))~c(fase[71:90],fase[91:96]), rho=0)


#Modelo de Cox

#SEXO

modsexo<-coxph(Surv(tempo,censura)~sexo, x=TRUE)

cox.zph(modsexo, transform="identity", global=TRUE)

plot(cox.zph(modsexo))

summary(modsexo)


#IDADE

modidade<-coxph(Surv(tempo,censura)~idade, x=TRUE)

cox.zph(modidade, transform="identity", global=TRUE)

plot(cox.zph(modidade))

summary(modidade)


#PLAQ

modplaq<-coxph(Surv(tempo,censura)~plaq, x=TRUE)

cox.zph(modplaq, transform="identity", global=TRUE)
```

```
plot(cox.zph(modplaq))

summary(modplaq)


#TEMPPLAQ

modtempplaq<-coxph(Surv(tempo,censura)~tempplaq, x=TRUE)

cox.zph(modtempplaq, transform="identity", global=TRUE)

plot(cox.zph(modtempplaq))

summary(modtempplaq)


#DEAG

moddeag<-coxph(Surv(tempo,censura)~deag, x=TRUE)

cox.zph(moddeag, transform="identity", global=TRUE)

plot(cox.zph(moddeag))


#TEMPDEAG

modtempdeag<-coxph(Surv(tempo,censura)~tempdeag, x=TRUE)

cox.zph(modtempdeag, transform="identity", global=TRUE)

plot(cox.zph(modtempdeag))

summary(modtempdeag)


#DECR

moddecr<-coxph(Surv(tempo,censura)~decr, x=TRUE)

cox.zph(moddecr, transform="identity", global=TRUE)
```

```
plot(cox.zph(moddecr))

summary(moddecr)



#TEMPDECR

modtempdecr<-coxph(Surv(tempo,censura)~tempdecr, x=TRUE)

cox.zph(modtempdecr, transform="identity", global=TRUE)

plot(cox.zph(modtempdecr))

summary(modtempdecr)


#FASE

modfase<-coxph(Surv(tempo,censura)~factor(fase), x=TRUE)

cox.zph(modfase, transform="identity", global=TRUE)

plot(cox.zph(modfase))

summary(modfase)


#Ajuste do Modelo

mod1<-
coxph(Surv(tempo,censura)~factor(sexo)+idade+plaq+tempplaq+deag+tempdeag+decr+temp
decr+factor(fase), x=TRUE)

cox.zph(mod1, transform="identity", global=TRUE)

plot(cox.zph(mod1))

summary(mod1)
```

```
mod2<-
coxph(Surv(tempo,censura)~idade+plaq+tempplaq+deag+tempdeag+decr+sexo+factor(fase),
x=TRUE)

cox.zph(mod2, transform="identity", global=TRUE)

plot(cox.zph(mod2))

summary(mod2)

mod3<-
coxph(Surv(tempo,censura)~idade+plaq+deag+tempdeag+sexo+tempplaq+factor(fase),
x=TRUE)

cox.zph(mod3, transform="identity", global=TRUE)

plot(cox.zph(mod3))

summary(mod3)


mod4<-coxph(Surv(tempo,censura)~idade+plaq+deag+tempdeag+tempplaq+factor(fase),
x=TRUE)

cox.zph(mod4, transform="identity", global=TRUE)

plot(cox.zph(mod4))

summary(mod4)


mod5<-coxph(Surv(tempo,censura)~plaq+deag+tempdeag+tempplaq+factor(fase), x=TRUE)

cox.zph(mod5, transform="identity", global=TRUE)

plot(cox.zph(mod5))

summary(mod5)
```

```
mod6<-coxph(Surv(tempo,censura)~plaq+tempdeag+tempplaq+factor(fase), x=TRUE)

cox.zph(mod6, transform="identity", global=TRUE)

plot(cox.zph(mod6))

summary(mod6)


mod7<-coxph(Surv(tempo,censura)~plaq+tempdeag+factor(fase), x=TRUE)

cox.zph(mod7, transform="identity", global=TRUE)

plot(cox.zph(mod7))

summary(mod7)


#Análise de Resíduos#


#Martingal

res.mart<-resid(mod7,type='martingal')

plot(res.mart,xlab='Índice',ylab='Resíduos Martingal',type='p')

abline(h=0)

identify(res.mart)




#Deviance

res.dev<-resid(mod7,type='deviance')
```

```
plot(res.dev,xlab='Índice',ylab='Resíduos Deviance',type='p')

abline(h=0)

identify(res.dev)


#INFLUÊNCIA GLOBAL


variancia<-mod7$var
hessi<-solve(variancia)


DadoN<-cbind(tempo,censura,plaq,tempdeag,fase)
n<-nrow(DadoN)
plaqcoef<-matrix(0,n,1)
tempdeagcoef<-matrix(0,n,1)
fasecoef1<-matrix(0,n,1)
fasecoef2<-matrix(0,n,1)
matrizvero<-matrix(0,n,1)


for(i in 1:n){
dad<-DadoN[-i, ]
tempoi<-dad[ ,1]
censurai<-dad[ ,2]
plaqi<-dad[ ,3]
tempdeagi<-dad[ ,4]
fasei<-dad[ ,5]


mod7i<-coxph(Surv(tempoi,censurai)~plaqi+tempdeagi+factor(fasei), x=TRUE)
plaqcoef[i]<-mod7i$coefficients[1]
tempdeagcoef[i]<-mod7i$coefficients[2]
fasecoef1[i]<-mod7i$coefficients[3]
fasecoef2[i]<-mod7i$coefficients[4]
```

```
matrizvero[i]<-mod7i$loglik[2]
}
plaqcoef
tempdeagcoef
fasecoef1
fasecoef2
matrizvero
paradel<-cbind(plaqcoef,tempdeagcoef,fasecoef1,fasecoef2)

DC<-(t(mod7$coefficients)-matrizpara[i])%*%hessi%*%(mod7$coefficients-matrizpara[i])
invhessi<-inv(hessi)
DC<-(t(mod7$coefficients)-matrizpara[i])%*%invhessi%*%(mod7$coefficients-
matrizpara[i])
Cookh<-matrix(0,n,1)
Cookv<-matrix(0,n,1)

for(j in 1:n){
Cookh[j]<-t(paradel[j,]-mod7$coefficients)%*%hessi%*%(paradel[j,]-mod7$coefficients)
Cookv[j]<-t(paradel[j,]-mod7$coefficients)%*%variancia%*%(paradel[j,]-
mod7$coefficients)
}
Cookh
Cookv

#Gráficos#
indice<-c(1:96)
LD<-2*(mod7$loglik[2]-matrizvero)
plot(indice,LD)
identify(indice,LD)
plot(indice,Cookh)
identify(indice,Cookh)
```

```
plot(indice,Cookv)
identify(indice,Cookv)

#Retirando as Observações Influentes

dad1<-dados[-33, ]

dad2<-dados[-34, ]

dad3<-dados[-81, ]

dad4<-dados[-c(33,34), ]

dad5<-dados[-c(33,81), ]

dad6<-dados[-c(34,81), ]

dad7<-dados[-c(33,34,81), ]

tempo1<-dad1$tempo

censura1<-dad1$status

plaq1<-dad1$plaq

tempdeag1<-dad1$tempdeag

fase1<-dad1$fase

mod71<-coxph(Surv(tempo1,censura1)~plaq1+tempdeag1+factor(fase1), x=TRUE)

cox.zph(mod71, transform="identity", global=TRUE)

plot(cox.zph(mod71))

summary(mod71)
```

```
tempo2<-dad2$tempo

censura2<-dad2$status

plaq2<-dad2$plaq

tempdeag2<-dad2$tempdeag

fase2<-dad2$fase


mod72<-coxph(Surv(tempo2,censura2)~plaq2+tempdeag2+factor(fase2), x=TRUE)

cox.zph(mod72, transform="identity", global=TRUE)

plot(cox.zph(mod72))

summary(mod72)


tempo3<-dad3$tempo

censura3<-dad3$status

plaq3<-dad3$plaq

tempdeag3<-dad3$tempdeag

fase3<-dad3$fase


mod73<-coxph(Surv(tempo3,censura3)~plaq3+tempdeag3+factor(fase3), x=TRUE)

cox.zph(mod73, transform="identity", global=TRUE)

plot(cox.zph(mod73))

summary(mod73)
```

```
tempo4<-dad4$tempo

censura4<-dad4$status

plaq4<-dad4$plaq

tempdeag4<-dad4$tempdeag

fase4<-dad4$fase


mod74<-coxph(Surv(tempo4,censura4)~plaq4+tempdeag4+factor(fase4), x=TRUE)

cox.zph(mod74, transform="identity", global=TRUE)

plot(cox.zph(mod74))

summary(mod74)


tempo5<-dad5$tempo

censura5<-dad5$status

plaq5<-dad5$plaq

tempdeag5<-dad5$tempdeag

fase5<-dad5$fase


mod75<-coxph(Surv(tempo5,censura5)~plaq5+tempdeag5+factor(fase5), x=TRUE)

cox.zph(mod75, transform="identity", global=TRUE)

plot(cox.zph(mod75))

summary(mod75)


tempo6<-dad6$tempo
```

```
censura6<-dad6$status

plaq6<-dad6$plaq

tempdeag6<-dad6$tempdeag

fase6<-dad6$fase


mod76<-coxph(Surv(tempo6,censura6)~plaq6+tempdeag6+factor(fase6), x=TRUE)

cox.zph(mod76, transform="identity", global=TRUE)

plot(cox.zph(mod76))

summary(mod76)


tempo7<-dad7$tempo

censura7<-dad7$status

plaq7<-dad7$plaq

tempdeag7<-dad7$tempdeag

fase7<-dad7$fase


mod77<-coxph(Surv(tempo7,censura7)~plaq7+tempdeag7+factor(fase7), x=TRUE)

cox.zph(mod77, transform="identity", global=TRUE)

plot(cox.zph(mod77))

summary(mod77)
```

6 REFERÊNCIAS BIBLIOGRÁFICAS

BRESLOW, N.E. (1972) *Discussion of Professor Cox's Paper.* Journal of the Royal Statistical Society.

COLOSIMO, E. A. (2001). *Análise de Sobrevivência Aplicada.* Universidade de São Paulo, Escola Superior de Agricultura Luiz de Queiroz, Piracicaba.

COOK, R.D. (1977). *Detection of Influential Observations in Linear Regression*. Technometrics.

COX, D.R. (1975). *Partial Likelihood.* Biometrika.

FACHINI, J. B. (2011). *Modelos de regressão com e sem fração de cura para dados bivariados em análise de sobrevivência*. Universidade de São Paulo, Escola Superior de Agricultura Luiz de Queiroz, Piracicaba.

FACHINI, J. B. (2006). *Análise de influência local nos modelos de riscos múltiplos*. Universidade de São Paulo, Escola Superior de Agricultura Luiz de Queiroz, Piracicaba.

GOMES, E. M. C (2007). *Análise de sensibilidade e resíduos em modelos de regressão com respostas bivariadas por meio de cópulas*. Universidade de São Paulo, Escola Superior de Agricultura Luiz de Queiroz, Piracicaba, 2007.

HOSMER, D.W., LEMESHOW, S. (1999). *Applied Survival Analysis.* John Wiley and Sons, New York.

PETO, R. (1972) Contribuição à discussão do artigo do D. R. Cox. Journal of the Royal Statistical Society.